Macromolecular crystallographic refinement

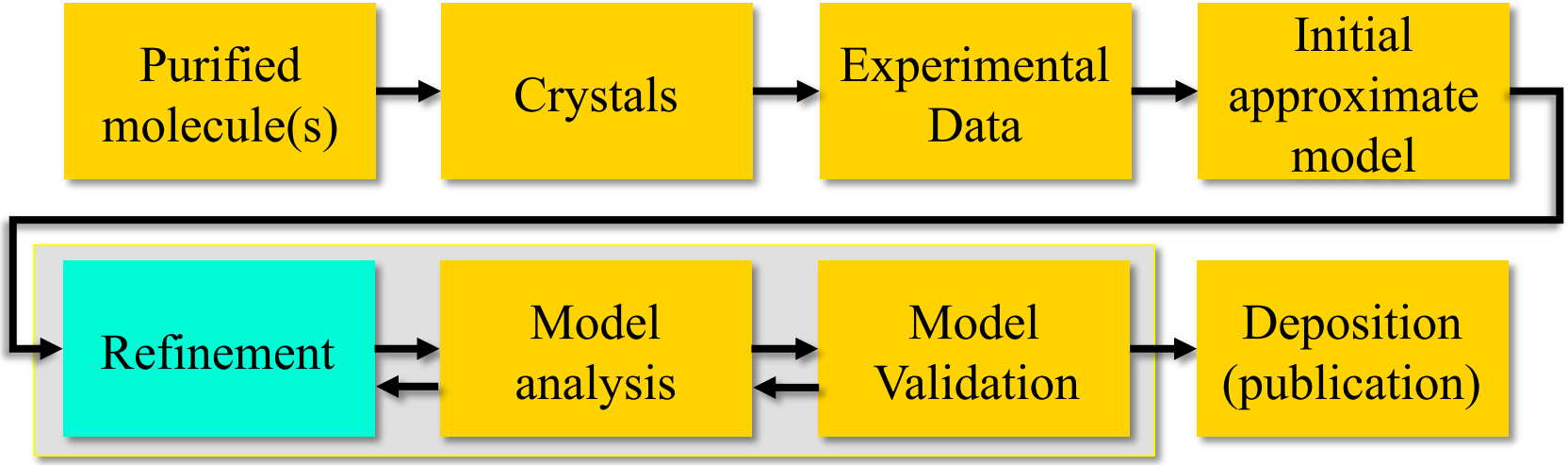
CCP4 School – Sao Carlos, Brazil, Nov 2018

(this presentation will be made available)

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

KING'S College LONDON
University of London

Crystallographic macromolecular refinement



Crystallographic refinement is an iterative process in which an initial structural model is progressively modified to produce an updated model which is more consistent with the experimental data and chemical knowledge.

Updated model…what does that mean?

You've got a starting model…(phase problem 'solved')

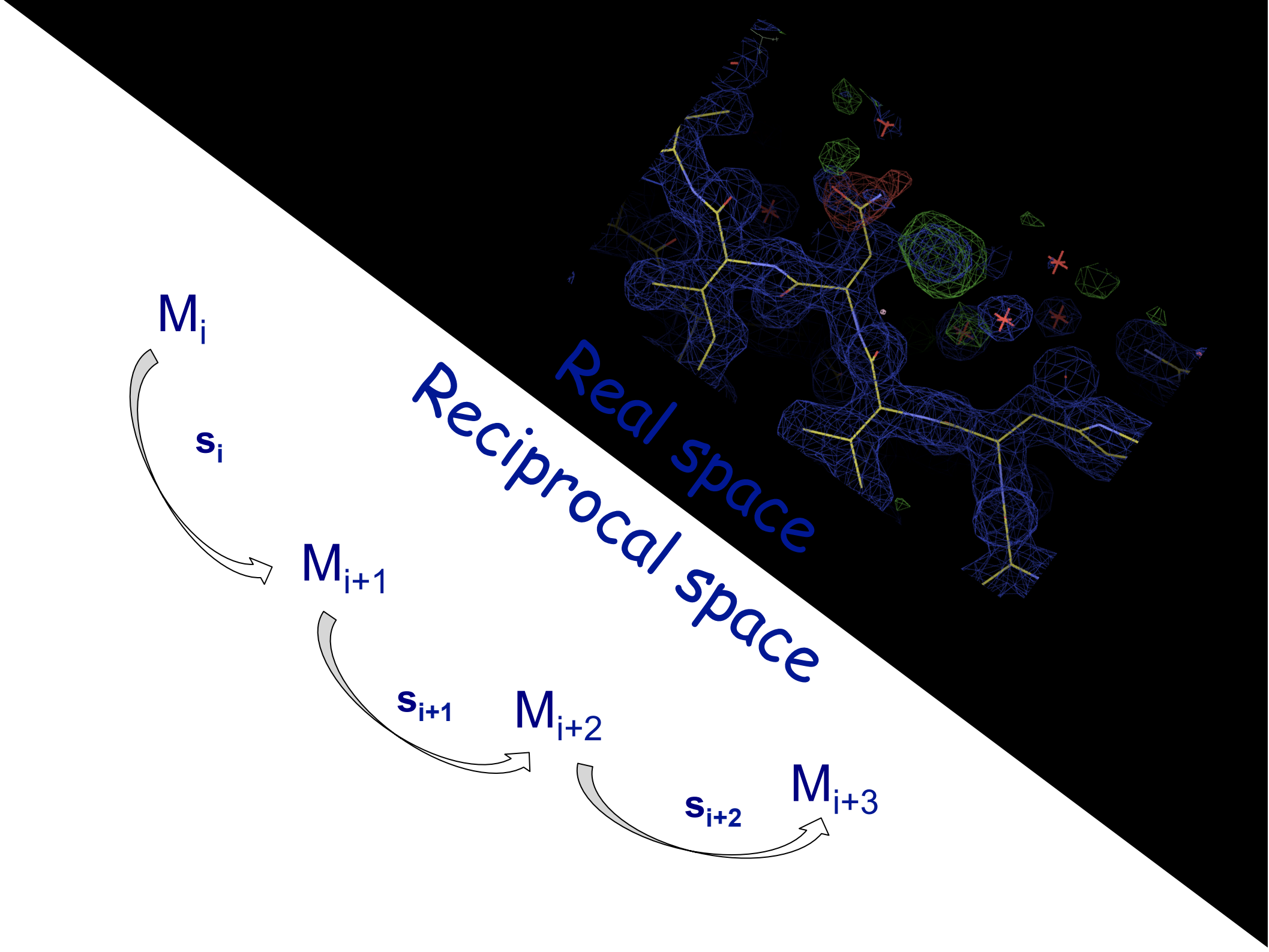
You want to improve it (typically optimise atom positions and thermal parmeters, add atoms - model completion) to satisfy what said before (experiment and chemistry).

$$R = \frac{\sum_h \left| \left|F_{\text{obs}}\right| - \left|F_{\text{calc}}\right| \right|}{\sum_h \left|F_{\text{obs}}\right|}.$$

Refinement is not only about low *R and* R_{free} factors.

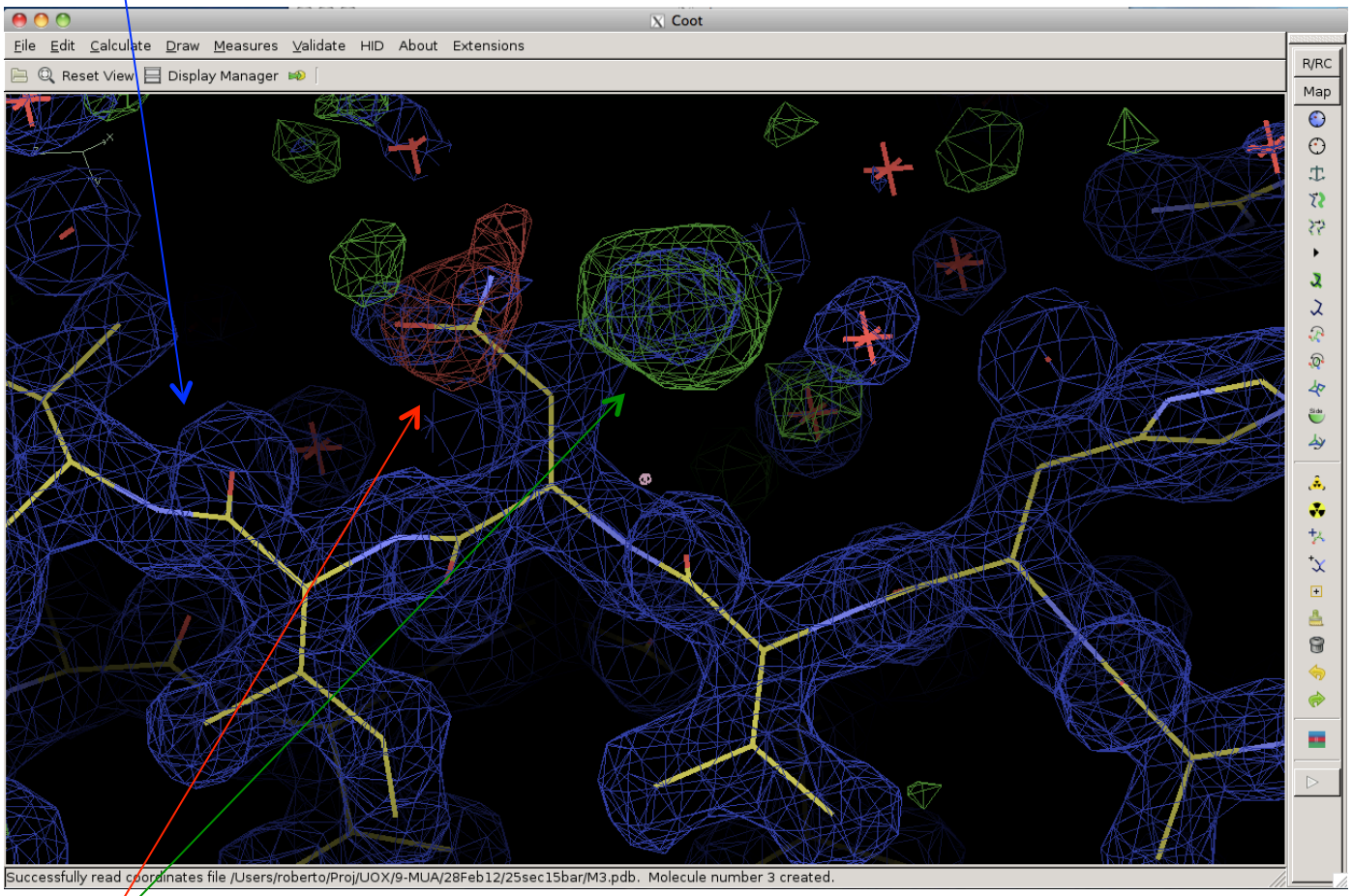
(This is not a good reason to be sloppy. Refinement is how you present your work to the world.)

Refinement is an iterative process that in practice is always terminated by the user.


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

$$\rho(xyz) = \frac{1}{V}\sum_{\text{hkl}} |F(hkl)| \exp[-2\pi i(hx+ky+lz) + i\varphi(hkl)]$$

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



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

Real space

Coot

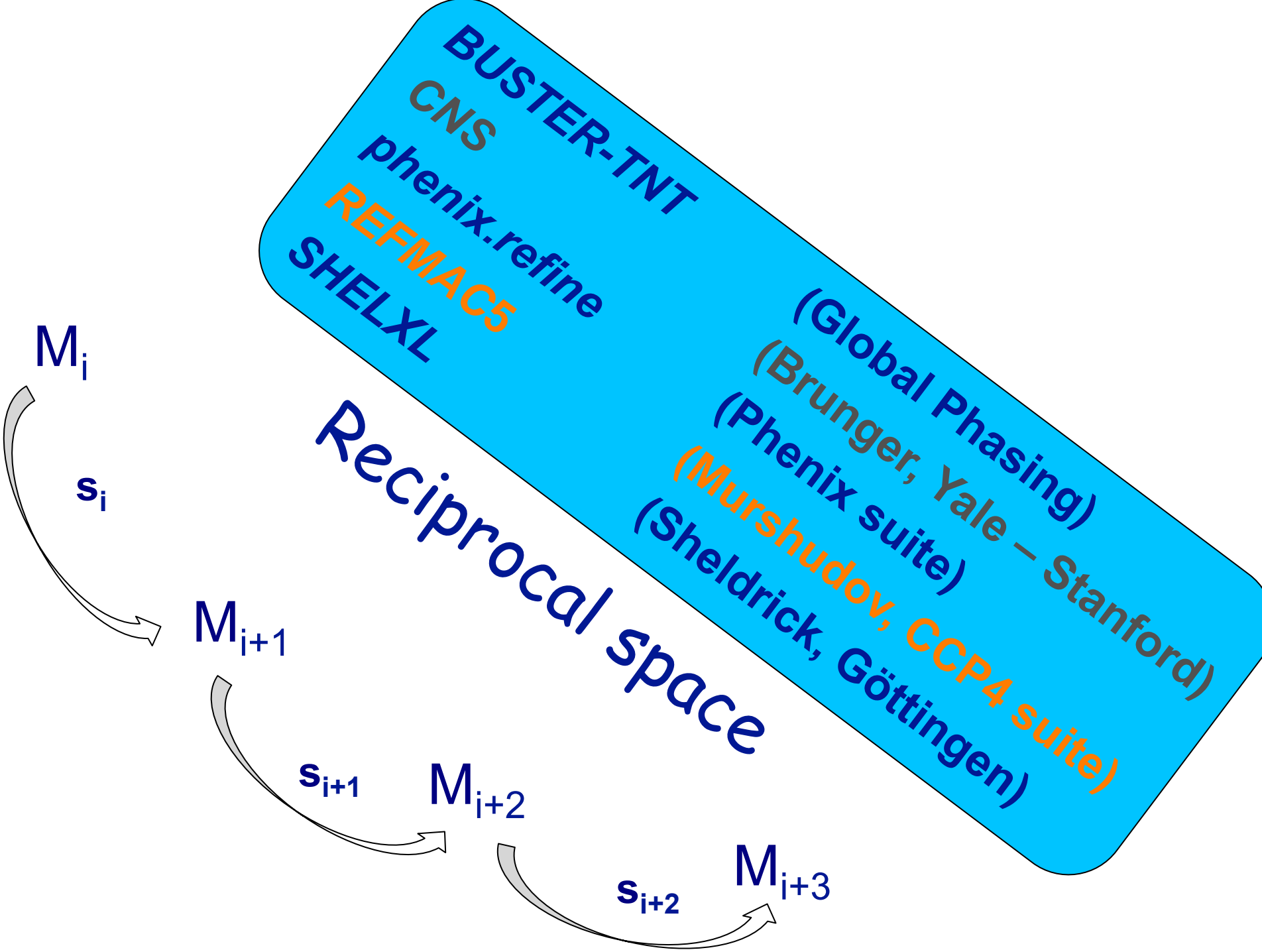
O

XtalView

(Emsley, Cambridge)

(Jones, Uppsala)

(McRae, San Diego)


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

REFMAC5

- Distributed as part of CCP4
- It is easy to use (CCP4*i* ➔ *CCP4i2*)
- Based on ML and Bayesian statistics
- Multiple tasks (model idealisation, rigid-body, jelly-body, restrained ML refinement, phased refinement)
- Automated twinned ML refinement
- Powerful and highly optimised minimisation algorithm (very fast)
- Extensive built-in dictionary (more than 11,000 library entries)
- Automatic X-ray/geometry weight estimation
- Flexible model parameterisation (iso-,aniso-, mixed-ADPs, TLS, bulk solvent, global and local NCS, occupancy)
- Low resolution tools (restraints to external structures and/or secondary structure ➔ Prosmart)
- Map sharpening
- Refinement engine of *ARP/wARP*, *BALBES*, *PDB_REDO*
- One-click viewing of results with *Coot*
- Extension to other techniques (cryoEM, ED, NMX,…)

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]

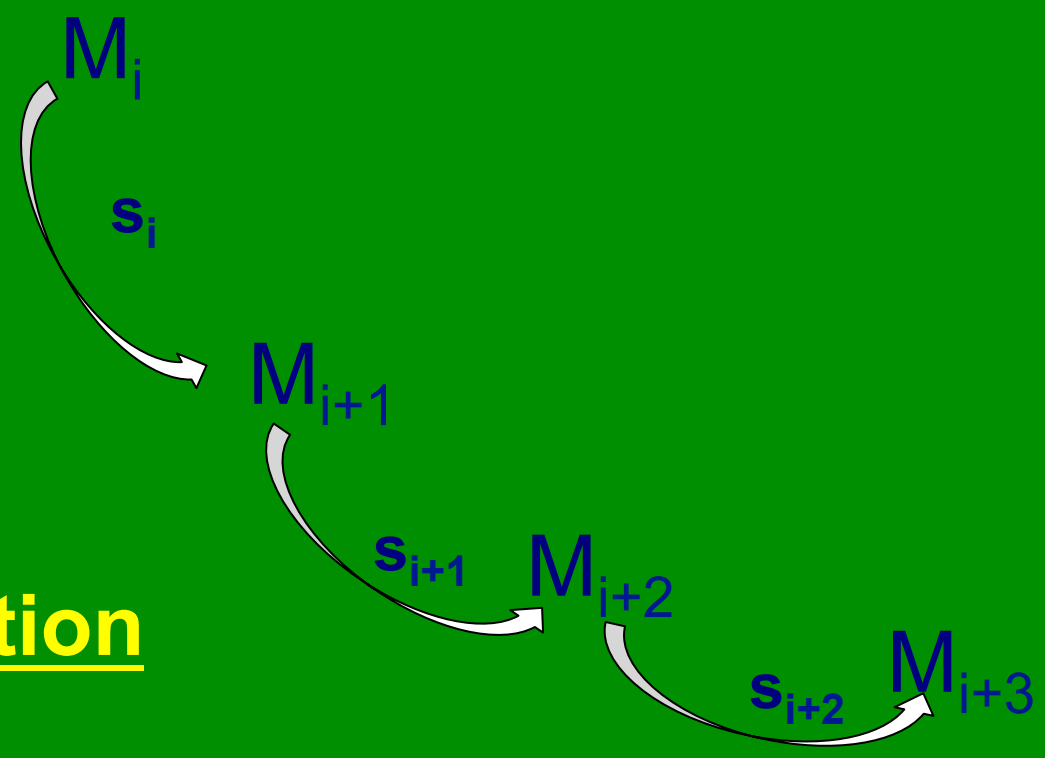
[a]Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5YW, England, [b]Biophysical Structural Chemistry, Leiden University, PO Box 9502, 2300 RA Leiden, The Netherlands, [c]Randall Division of Cell and Molecular Biophysics, New Hunt's House, King's College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

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

This paper describes various components of the macromolecular crystallographic refinement program *REFMAC*5, which is distributed as part of the *CCP*4 suite. *REFMAC*5 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, *REFMAC*5 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. *REFMAC*5 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. *REFMAC*5 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

Key aspects of (reciprocal space) refinement



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

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Introduction to macromolecular refinement

Dale. E. Tronrud

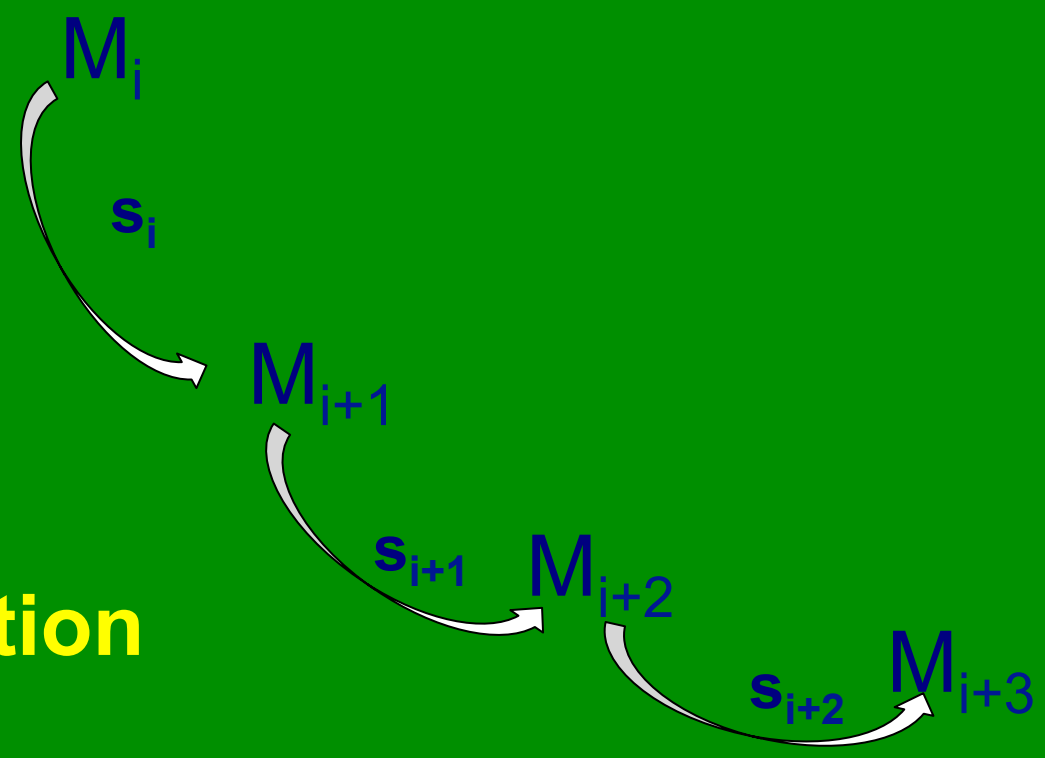
Howard Hughes Medical Institute and Institute of Molecular Biology, University of Oregon, Eugene, OR 97403, USA

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

Received 5 April 2004
Accepted 21 September 2004

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.

Key aspects of (reciprocal space) refinement



- <u>Objective function</u>
- Method of optimization
- Model parametrization
- Prior knowledge

For example, one could minimise a purely diffraction-based function (least-squares function)

$$f_{X-ray} = \sum_i w_i \left(\left| F_o \right|_i - \left| F_c \right| \right)^2$$

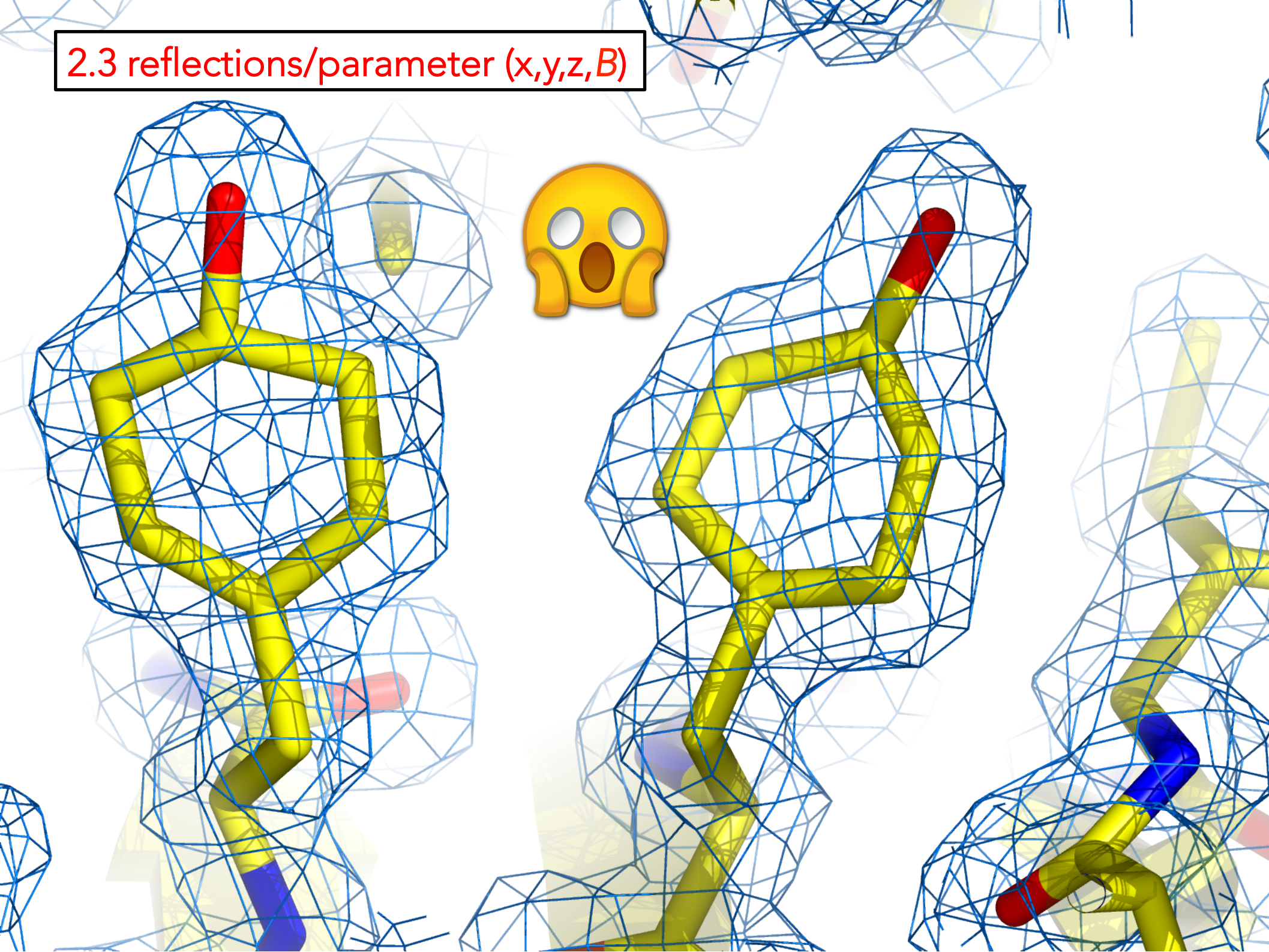
Macromolecular crystallography

In macromolecular crystallography the typically limited resolution of X-ray data combined with the size of the molecules under investigation results in an unfavorable data/parameters ratio.

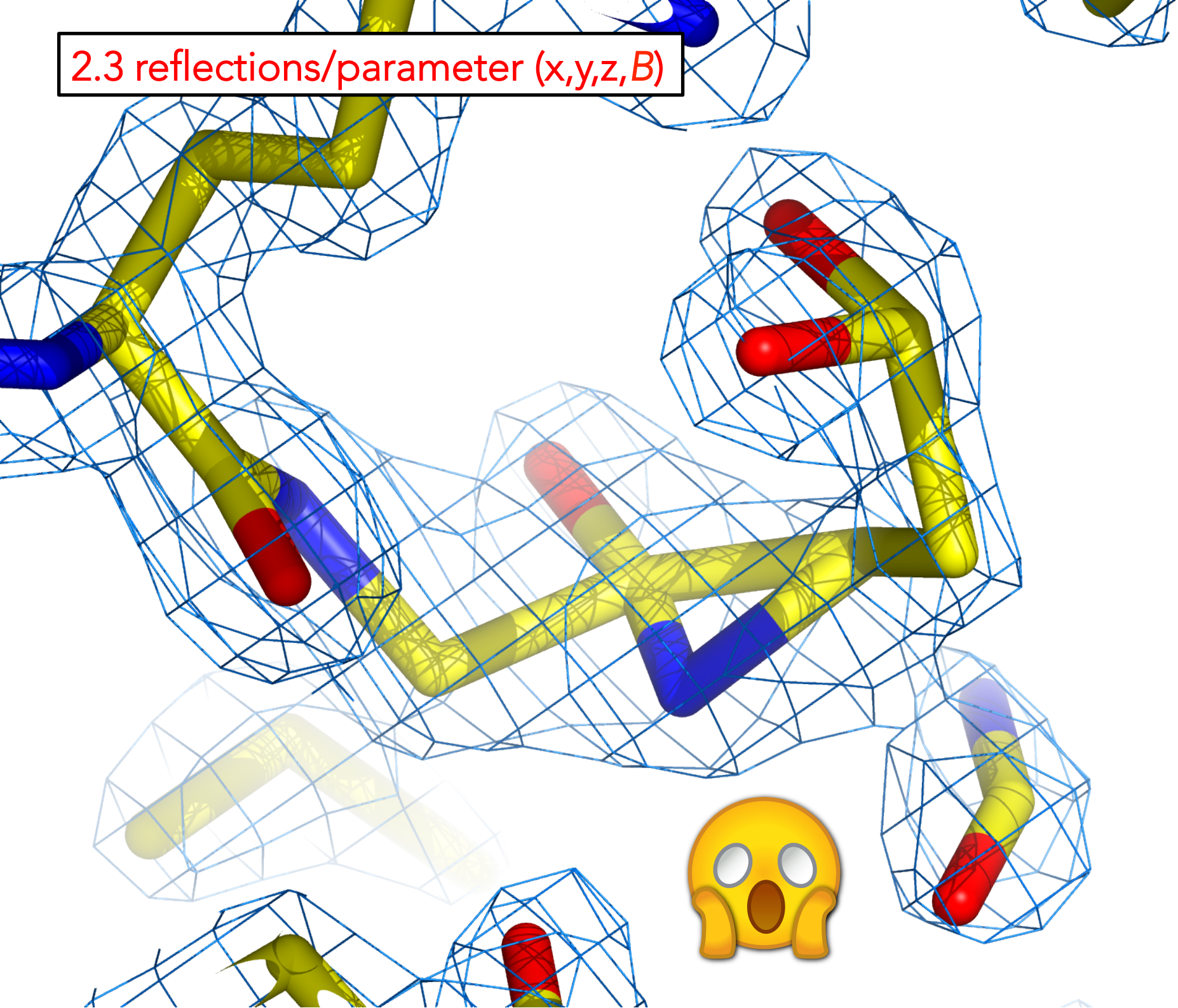
1.8 Å / 164 aa / 1540 non-H atoms / 14217 reflections
≈2.3 reflections/parameter (x,y,z,*B*)
≈1.0 reflections/parameter (x,y,z,*Us*)
≈100 for small molecules

Macromolecular refinement against solely X-ray data leads to severe model distortions reflecting unreasonable/impossible chemistry.

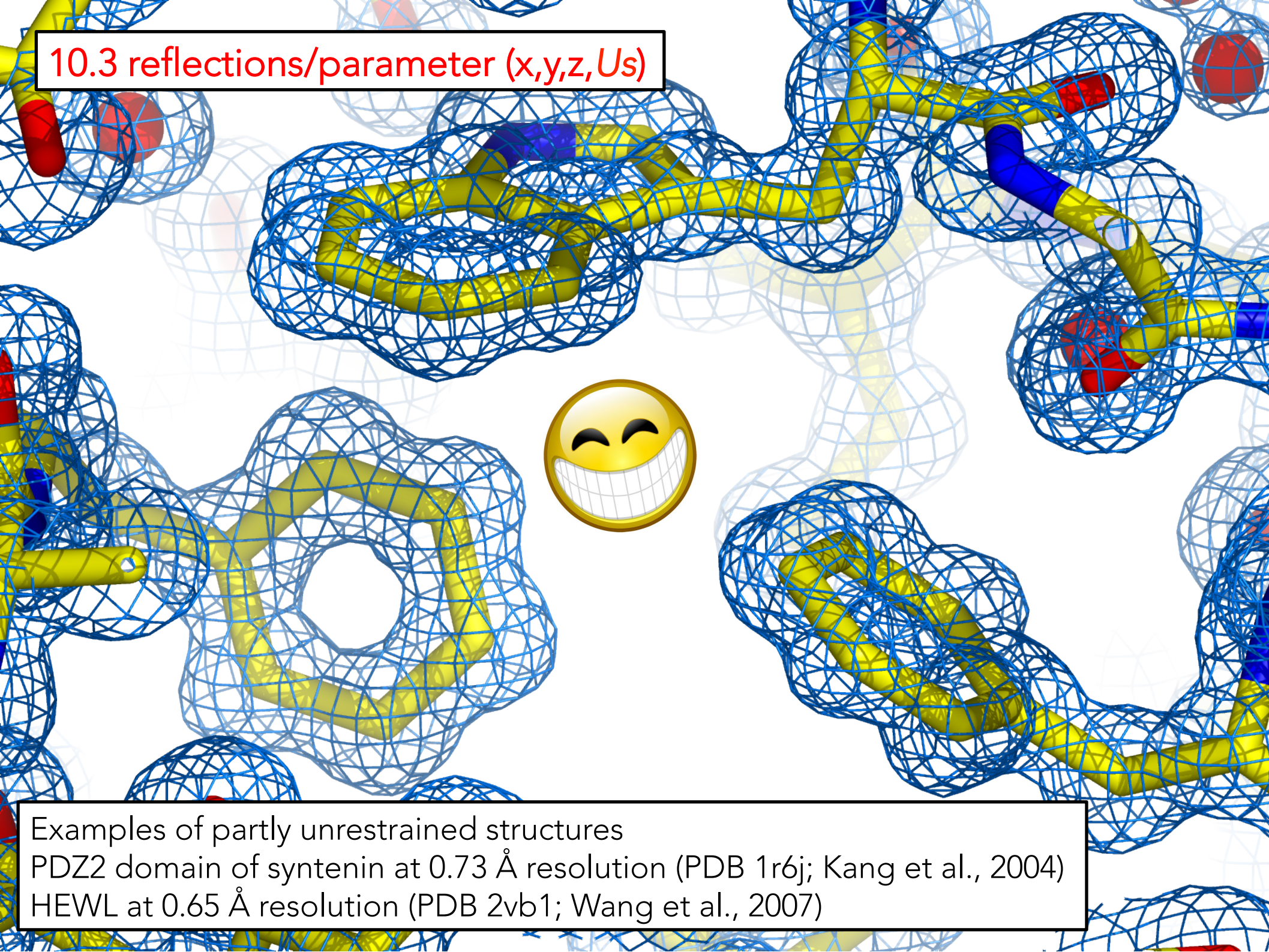
2.3 reflections/parameter (x,y,z,*B*)



2.3 reflections/parameter (x,y,z,*B*)

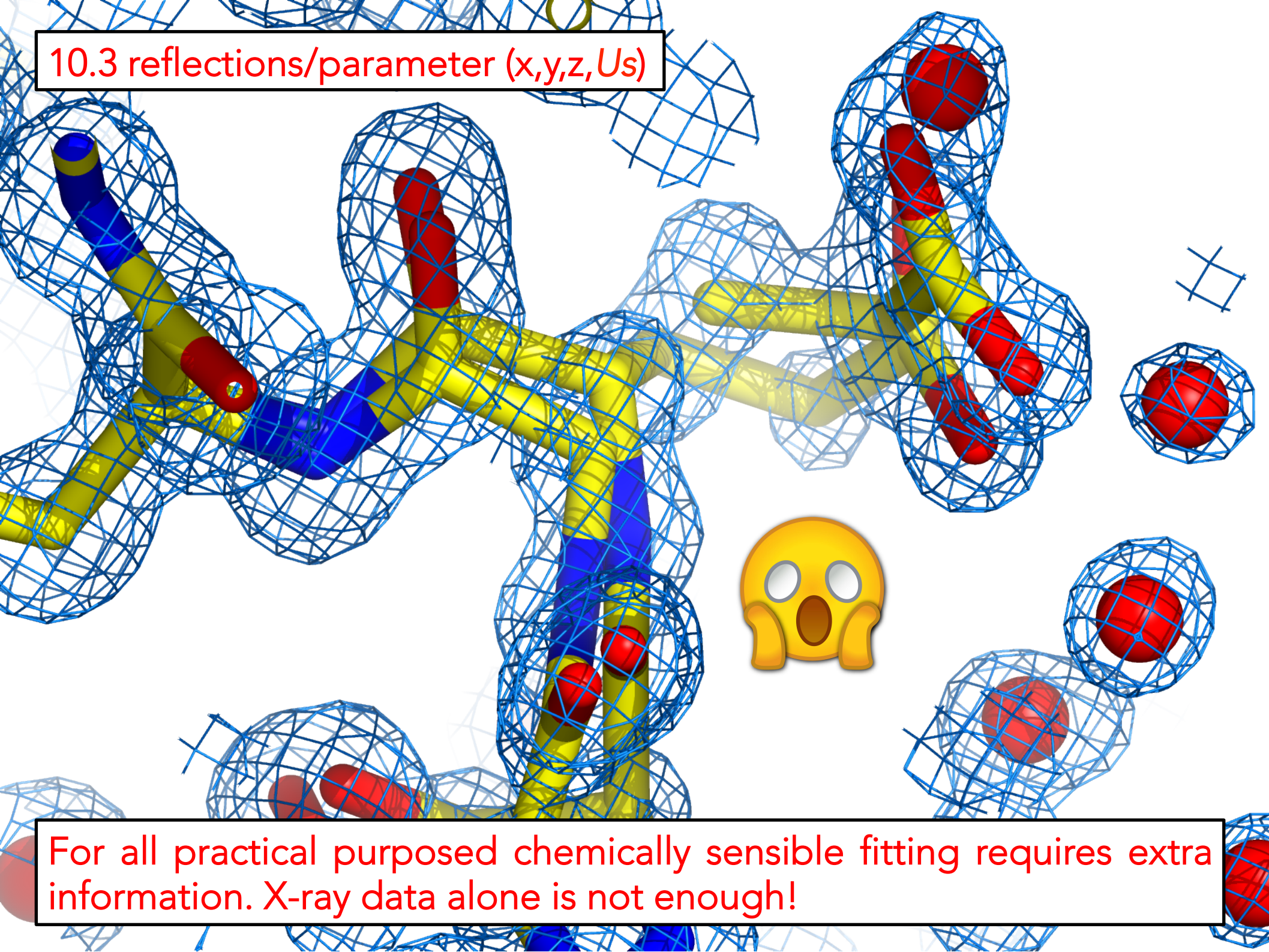


10.3 reflections/parameter (x,y,z,*Us*)



Examples of partly unrestrained structures
PDZ2 domain of syntenin at 0.73 Å resolution (PDB 1r6j; Kang et al., 2004)
HEWL at 0.65 Å resolution (PDB 2vb1; Wang et al., 2007)

10.3 reflections/parameter (x,y,z,*Us*)



For all practical purposed chemically sensible fitting requires extra information. X-ray data alone is not enough!

Subsidiary conditions / restraints

Something must be done to obtain <u>chemically sensible structural models</u>.

Acta Cryst. (1963). **16**, 1091

Least-Squares Refinement with Subsidiary Conditions

By Jürg Waser

Gates and Crellin Laboratories of Chemistry, California, Institute of Technology, Pasadena, California, U.S.A.*

(*Received* 18 *January* 1963)

A method of least-squares refinement is described in which the subsidiary conditions are treated like observational equations. The advantages of the method are its generality, its adaptability to machine computing, and the possibility of relaxing the subsidiary conditions to any desired degree by appropriate changes in the weighting. In suitable cases the method extends the range for which least-squares refinements converge to the correct solution.

$$f = \sum_i w_i \left(|F_o|_i - |F_c| \right)^2 + \sum_l w_l \left(p_{\text{model},l} - p_{\text{target},l} \right)^2$$

Restraints ≠ Constraints

<u>Restraints</u> are treated like observations and have a probability distribution

<u>Constraints</u> describe a mathematical condition (q1+q2=1, rigid-bodies,..)

$$f_{total} = wf_{X-ray} +$$

$$f_{bonds} + f_{angles} + f_{dihedrals} + f_{planarity} + f_{non-bonded} + f_{chirality} +$$

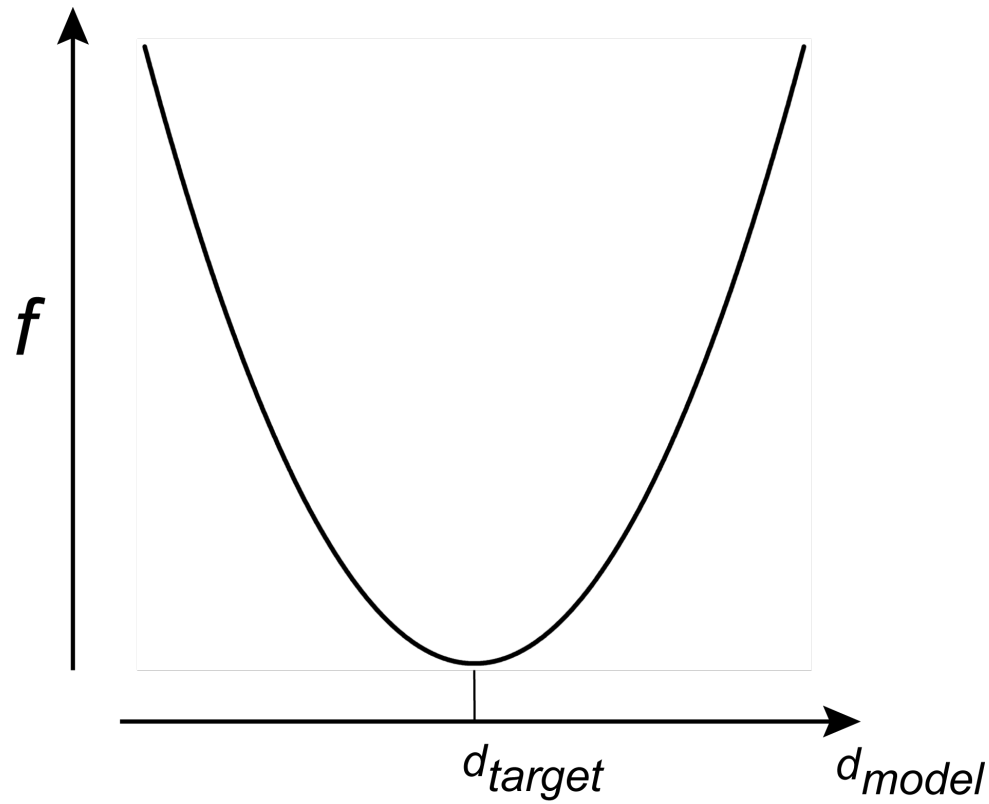
$$f_{NCS} + f_{reference} + f_{\substack{\text{secondary}\\ \text{structure}}} + \ldots\ldots$$

Some examples of restraints

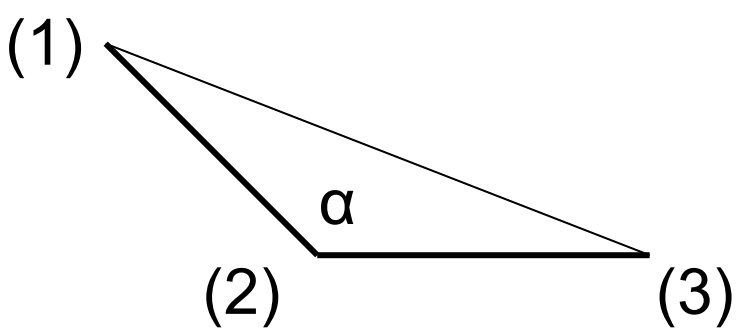
Bonds/Angles

$$f_{bonds} = \sum_{bonds} \frac{1}{\sigma_{bond}^2}\left(d_{\text{model}} - d_{\text{target}}\right)^2$$

$$f_{angles} = \sum_{angles} \frac{1}{\sigma_{angle}^2}\left(\alpha_{\text{model}} - \alpha_{\text{target}}\right)^2$$



Alternatively, one can restrain 1-3 distances:



$$f_{\text{1-3 distances}} = \sum_{\text{1-3}} \frac{1}{\sigma_{1-3}^2}\left(d_{\text{1-3 model}} - d_{\text{1-3 target}}\right)^2$$

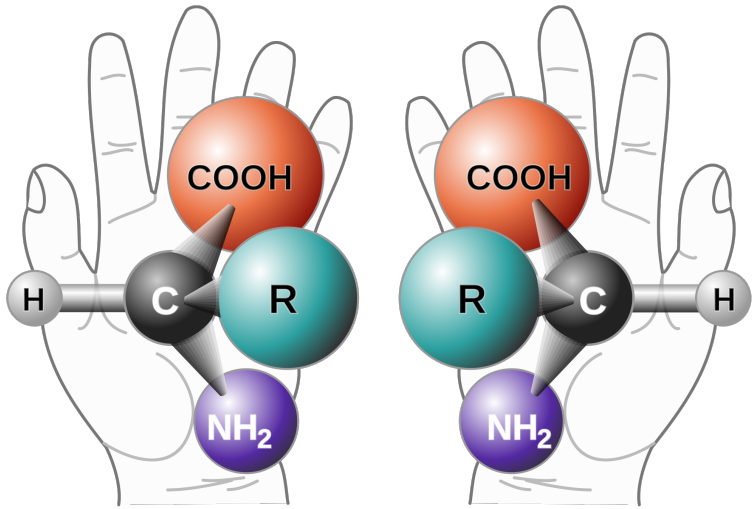
Some examples of restraints

Chirality

$$V = \left(\mathbf{r}_{\mathrm{N}} - \mathbf{r}_{\mathrm{CA}}\right) \bullet \left[\left(\mathbf{r}_{\mathrm{C}} - \mathbf{r}_{\mathrm{CA}}\right) \times \left(\mathbf{r}_{\mathrm{CB}} - \mathbf{r}_{\mathrm{CA}}\right)\right]$$

$$V_D = -V_L$$

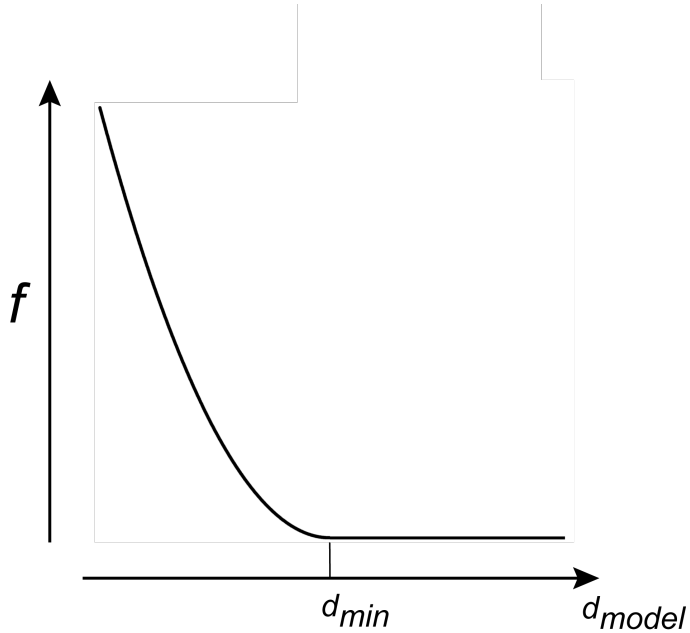
$$f_{chirals} = \sum_{chirals} \frac{1}{\sigma_{chiral}^2} \left(V_{\mathrm{model}} - V_{\mathrm{target}}\right)^2$$



Non-bonded

$$f_{nb} = \sum_{nb} \frac{1}{\sigma_{nb}^2} \left(d_{\mathrm{model}} - d_{\mathrm{min}}\right)^2$$

$$\text{if } \left(\mathrm{d}_{\mathrm{model}} < \mathrm{d}_{\mathrm{min}}\right)$$



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 is affected by the disease 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 do 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

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

$$L(O;M)$$

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

[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)

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

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

Objective (target) function

2. Target functions in *REFMAC5*

As in all other refinement programs, the target function minimized in *REFMAC5* has two components: a component utilizing geometry (or prior knowledge) and a component utilizing experimental X-ray knowledge,

$$f_{\text{total}} = f_{\text{geom}} + w f_{\text{xray}}, \tag{1}$$

where f_{total} is the total target function to be minimized, consisting of functions controlling the geometry of the model and the fit of the model parameters to the experimental data, and w is a weight between the relative contributions of these two components. In macromolecular crystallography, the weight is traditionally selected by trial and error. *REFMAC5* offers automatic weighting, which is based on the fact that both components are the natural logarithm of a probability distribution. However, this 'automatic' weight may lead to unrea-

$$f_{\text{total}} = -\log[P_{\text{posterior}}(\text{model; obs})]$$
$$f_{\text{geom}} = -\log[P_{\text{prior}}(\text{model})]$$
$$f_{\text{xray}} = -\log[P_{\text{likelihood}}(\text{obs; model})].$$

Likelihood (1)

2.1. X-ray component

The X-ray likelihood target functions used in *REFMAC5* are based on a general multivariate probability distribution of E observations given M model structure factors. This function is derived from a multivariate complex Gaussian distribution of $N = E + M$ structure factors for acentric reflections and from a multivariate real Gaussian distribution for centric reflections and has the following form:

$$P = \begin{cases} \dfrac{|\mathrm{C}_M| \prod_{i=1}^{E} |F_i|}{\pi^E |\mathrm{C}_N|} \int_0^{2\pi} \cdots \int_0^{2\pi} P_{\mathrm{pr}}(\boldsymbol{\alpha}) \\ \times \exp\left[-\sum_{i,j=1}^{N} F_i (a_{i,j} - c_{i-E,j-E}) F_j \right] \mathrm{d}\boldsymbol{\alpha} & \text{acentric} \\ \left[\dfrac{|\mathrm{C}_M|}{(2\pi)^E |\mathrm{C}_N|} \right]^{1/2} \sum_{\substack{\alpha_1=\alpha_{1,1} \\ \alpha_1=\alpha_{1,2}}} \cdots \sum_{\substack{\alpha_E=\alpha_{E,1} \\ \alpha_E=\alpha_{E,2}}} P_{\mathrm{pr}}(\boldsymbol{\alpha}) \\ \times \exp\left[-\dfrac{1}{2} \sum_{i,j=1}^{N} F_i (a_{i,j} - c_{i-E,j-E}) F_j \right] & \text{centric} \end{cases}, \quad (3)$$

where $P = P(|F_1|, \ldots, |F_E|; F_{E+1}, \ldots, F_N)$, $F_i = |F_i| \exp(\iota \alpha_i)$, $|F_1|, \ldots, |F_E|$ denote the observed amplitudes, $F_{E+1}, \ldots, F_N$ are the model structure factors, C_N is the covariance matrix with the elements of its inverse denoted by a_{ij}, C_M is the bottom right square submatrix of C_N of dimension M with the elements of its inverse denoted by c_{ij}. We define $c_{ij} = 0$ for $i \leq 0$ or $j \leq 0$. $|\mathrm{C}_N|$ and $|\mathrm{C}_M|$ are the determinants of matrices C_N and C_M, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_E)$ is the vector of the unknown phases of the observations that need to be integrated and $P_{\mathrm{pr}}(\boldsymbol{\alpha})$ is a probability distribution expressing any prior knowledge about the phases.

Likelihood (2)

In the simplest case of one observation, one model and no prior knowledge about phases, the integral in (3) can be evaluated analytically. In this case, the function follows a Rice distribution (Bricogne & Irwin, 1996), which is a non-central χ^2 distribution of $|F_o|^2/\Sigma$ and $|F_o|^2/2\Sigma$ with non-centrality parameters $D^2|F_c|^2/\Sigma$ and $D^2|F_o|^2/2\Sigma$ with one and two degrees of freedom for centric and acentric reflections, respectively (Stuart & Ord, 2009),

$$P(|F_o|; F_c) = \begin{cases} \dfrac{2|F_o|}{\Sigma}\exp\left(-\dfrac{|F_o|^2 + D^2|F_c|^2}{\Sigma}\right) \times I_0\left(2\dfrac{|F_o|D|F_c|}{\Sigma}\right) & \text{acentric} \\ \left(\dfrac{2}{\pi\Sigma}\right)^{1/2}\exp\left(-\dfrac{|F_o|^2 + D^2|F_c|^2}{2\Sigma}\right) \times \cosh\left(\dfrac{|F_o|D|F_c|}{\Sigma}\right) & \text{centric} \end{cases}, \quad (4)$$

where D in its simplest interpretation is $\langle\cos(\Delta xs)\rangle$, a Luzzati error parameter (Luzzati, 1952) expressing errors in the positional parameters of the model, F_c is the model structure factor, $|F_o|$ is the observed amplitude of the structure factor and Σ is the uncertainty or the second central moment of the distribution. Both Σ and D enter the equation as part of the covariance matrices C_N and C_M from (3). Σ is a function of the multiplicity of the Miller indices (ε factor), experimental uncertainties (σ_o), model completeness and model errors. For simplicity, the following parameterization is used:

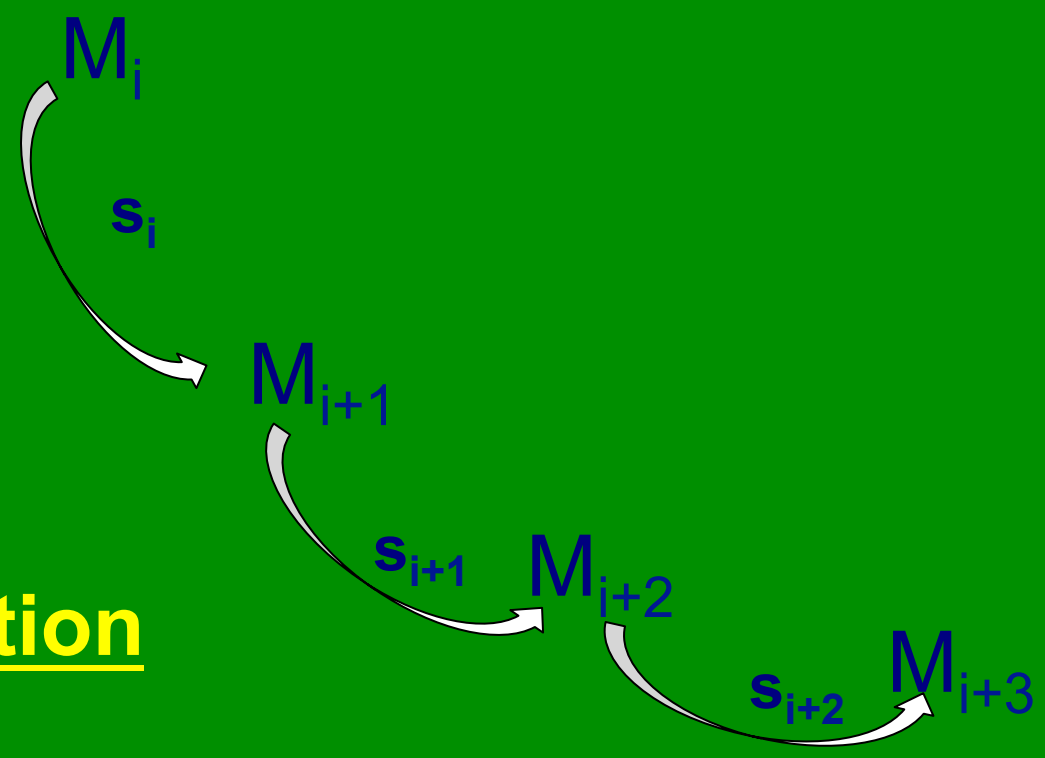
$$\Sigma = \begin{cases} 2\sigma_o^2 + \varepsilon\Sigma_{mod} & \text{acentric} \\ \sigma_o^2 + \varepsilon\Sigma_{mod} & \text{centric} \end{cases}. \quad (5)$$

The current version of *REFMAC5* estimates D and Σ_{mod} in resolution bins. Working reflections are used for estimation of D and free reflections are used for Σ_{mod} estimation. Although this simple parameterization works in many cases, it may give misleading results for data from crystals with pseudo translation, OD disorder or modulated crystals in general. Currently, there is no satisfactory implementation of the error model to account for these cases.

Summary object function

- The only real parameter the user can play with is the weight factor between X-ray and geom components of the objective function.
- Refemac5, Buster, phenix.refine all use ML functions. ShelxL uses LS.

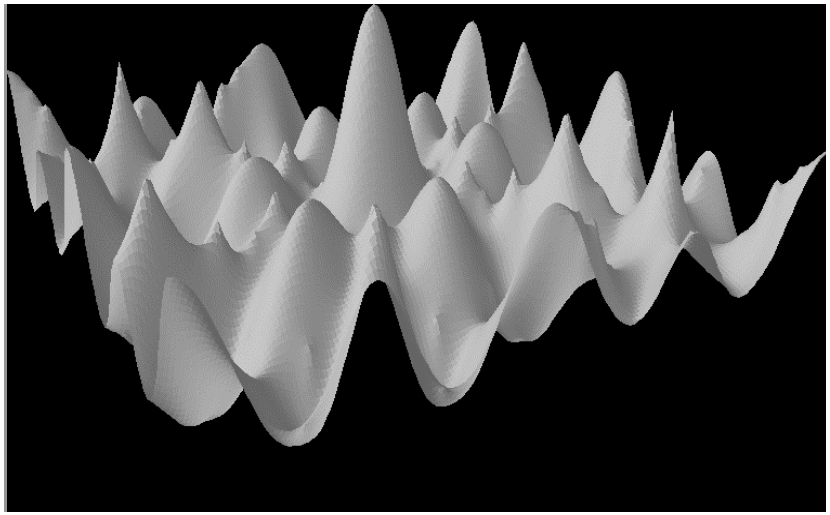
Key aspects of (reciprocal space) refinement



- **Objective function**
- **<u>Method of optimization</u>**
- **Model parametrization**
- **Prior knowledge**

Convergence

- Landscape of a refinement function is very complex

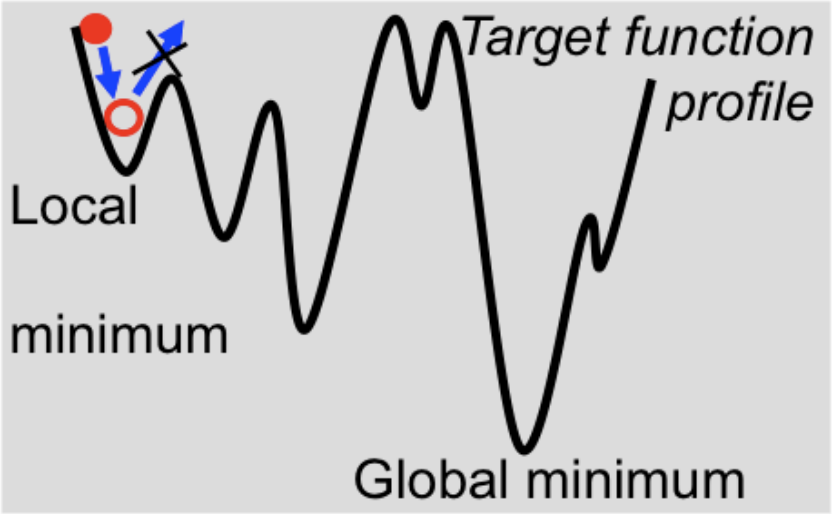


Picture stolen from Dale Tronrud

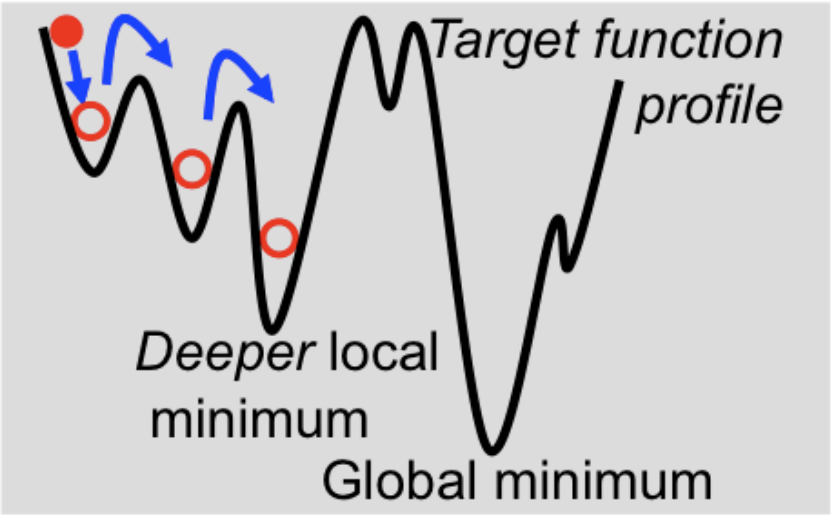
- Refinement programs have very small convergence radii compared to the size of the function profile. Depending where you start, the refinement engine will bring the structure to one of the closest local minimum

Refinement target optimization methods (from Pavel)

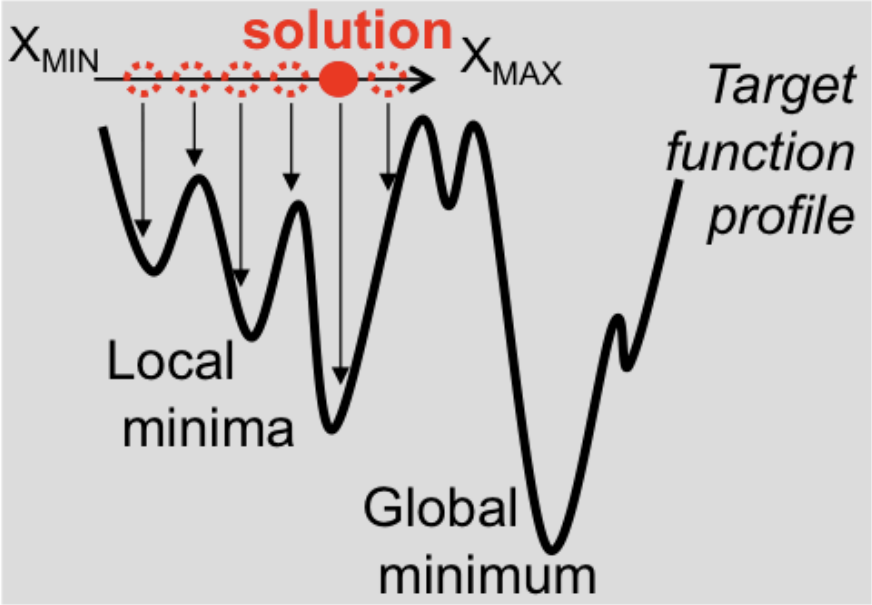
- **Gradient-driven minimization**



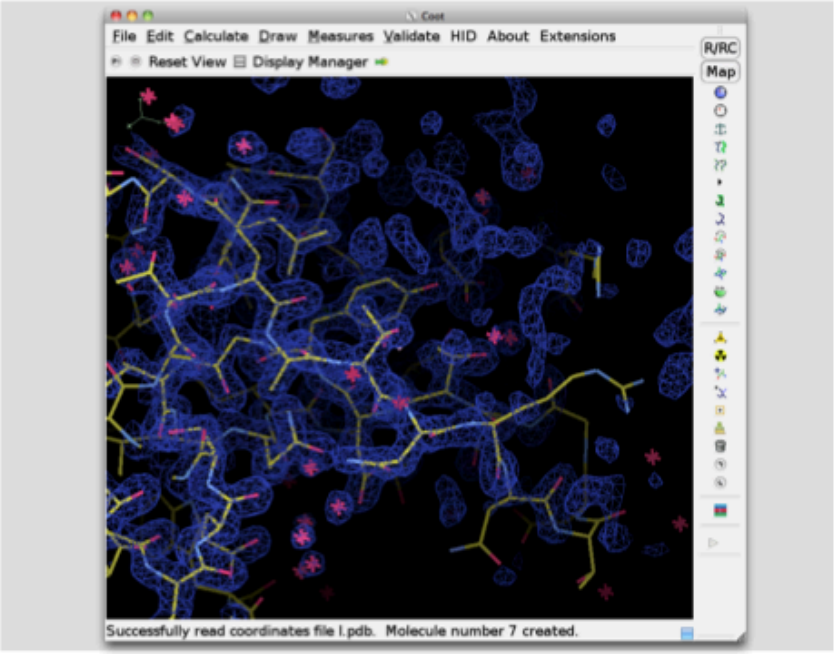
- **Simulated annealing (SA)**



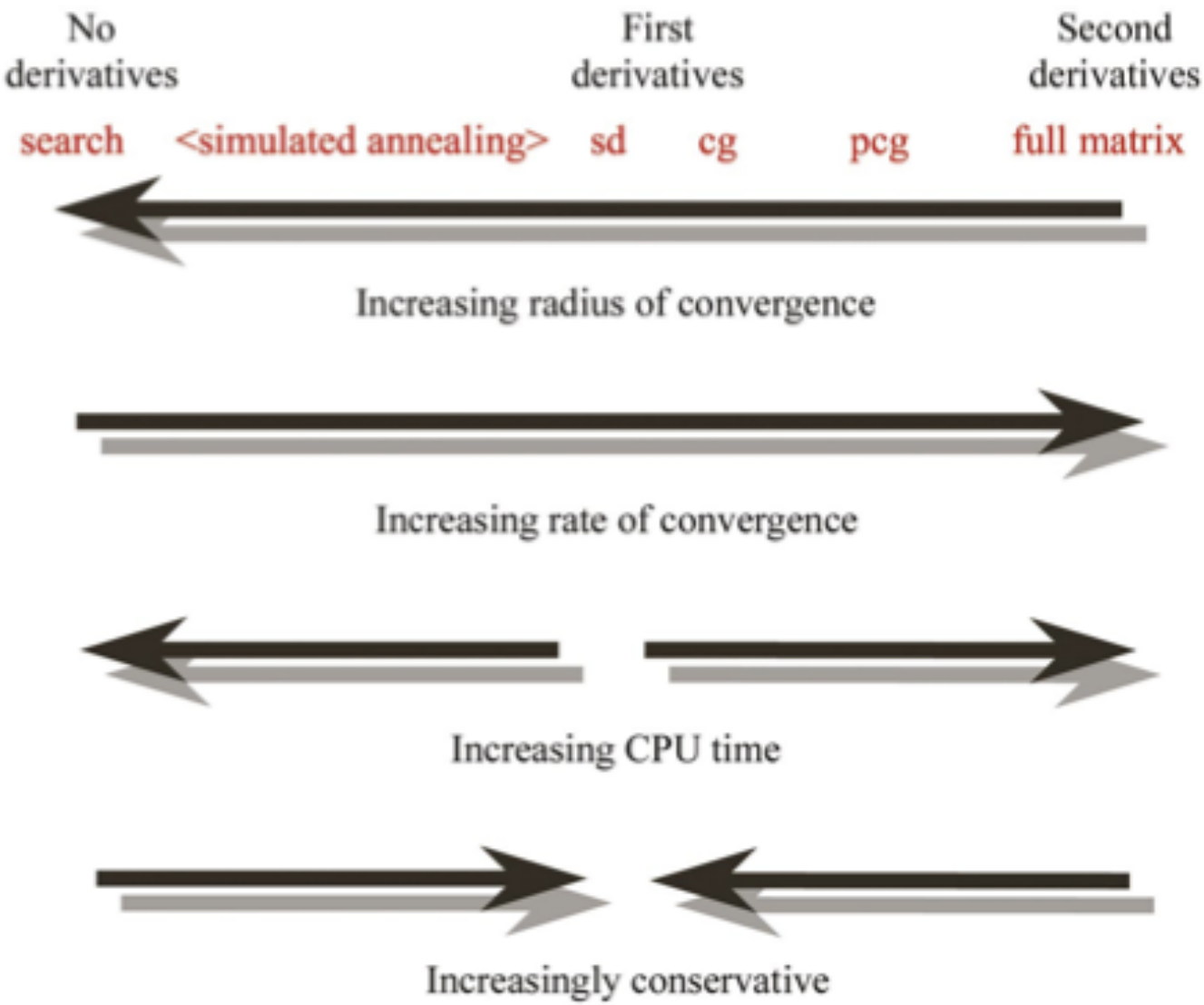
- **Grid search** (Sample parameter space within known range [X_{MIN}, X_{MAX}])



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



Overview optimisation methods



Macromolecules

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

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

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

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

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

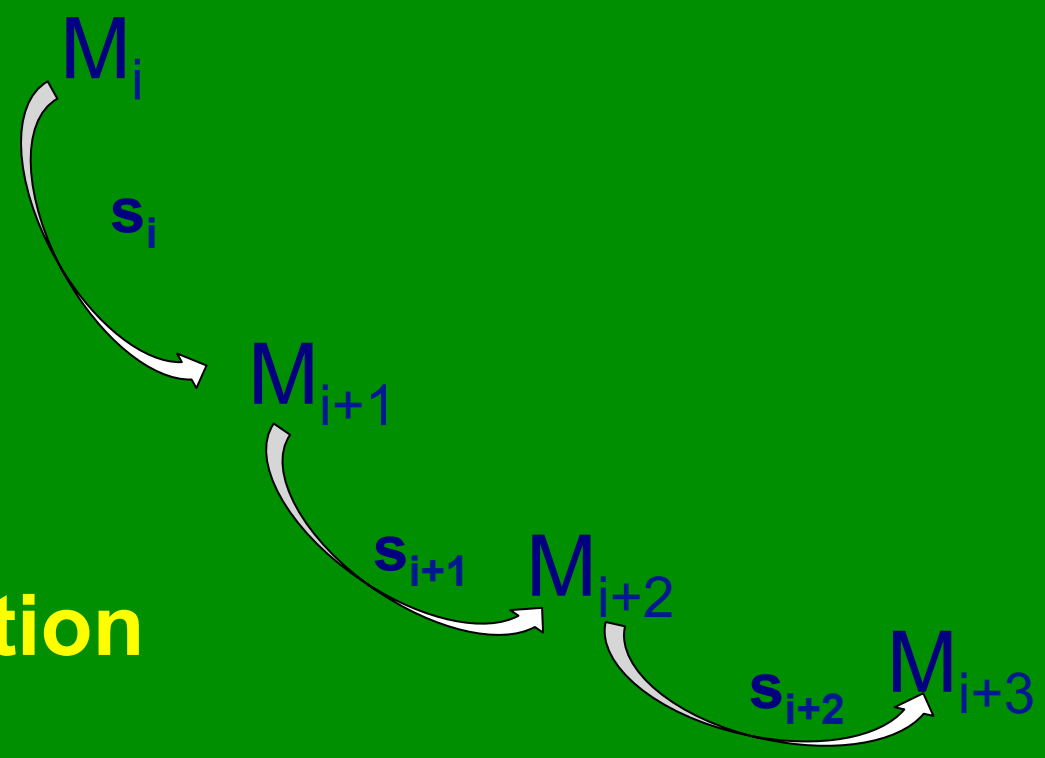
Received 13 June 2003
Accepted 21 August 2003

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 *REFMAC*5 of the *CCP*4 suite. The Fisher's matrix is used in its sparse approximation. Tests indicate that the algorithm performs satisfactorily.

Summary minimization

- As user nothing to change.
- Refmac5 uses a sparse matrix.
- Computational optimisation enables fast calculations thus allowing to take advantage of an increased rate of convergence without time overhead.

Key aspects of (reciprocal space) refinement



- **Objective function**
- **Method of optimization**
- **<u>Model parametrization</u>**
- **Prior knowledge**

How is the crystal content parametrised?

Atomic parameters

- Coordinates (x,y,z)
- ADPs (ISO or ANISO)
- Occupancies

Non-Atomic parameters

- Bulk solvent
- Crystal-specific
  - Anisotropy
  - Twinning

Non-atomic parameters

$$F_{MODEL} = k_{OVERALL} e^{-sU_{CRYSTAL} s^t} \left(F_{CALC\_ATOMS} + k_{SOL} e^{-\frac{B_{SOL} s^2}{4}} F_{MASK} \right)$$

Anisotropy **Bulk-solvent contribution**

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)

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

Bulk solvent

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

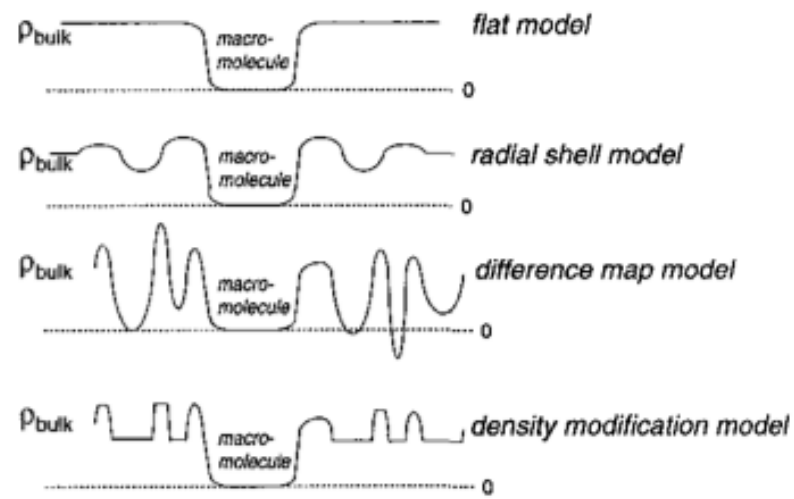
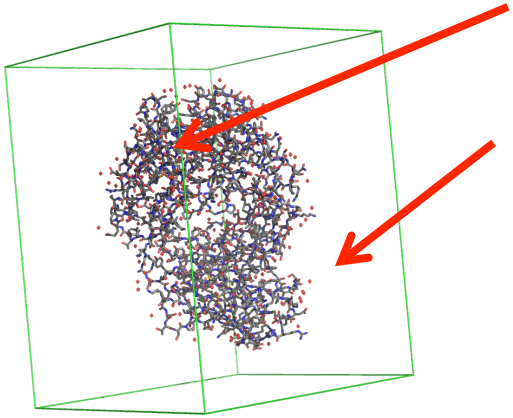


Figure 1. Schematic illustration for the 4 solvent models that were tested; flat model, radial shell model, difference map model and density modification model. The models are described in detail in the text.



Macromolecule region

Solvent region

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

Twinning

NEEDS A SEPARATE TALK

TOTALLY AUTOMATED IN REFMAC5

Atomic parameters

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)

<u>Atomic model</u> 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").

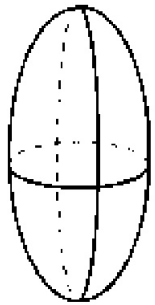
Atomic Displacement Parameters

For the purposes of discussion, it is convenient to consider four separate (and in general anisotropic) contributions to the total atomic displacement parameter,

$$U = U_{\text{crystal}} + U_{\text{TLS}} + U_{\text{internal}} + U_{\text{atom}}. \qquad (1)$$

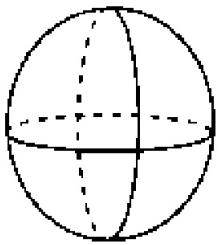
U_{crystal} represents the overall anisotropy of the crystal and is a single anisotropic displacement parameter applied to the entire contents of the unit cell; as such it obeys the symmetry of the crystal space group when refined against merged data. Inclusion of such anisotropic scaling is known to give improvements in crystallographic R and free R factors of up to several percent and improved behaviour of refinement (Sheriff & Hendrickson, 1987; Murshudov *et al.*, 1998). U_{TLS} represents translations and librations of pseudo-rigid bodies within the asymmetric unit of the crystal. These bodies may be whole molecules or identifiable molecular subunits. Next, U_{internal} includes various kinds of intramolecular collective motions, such as libration about particular torsion angles or internal normal modes of a molecule. Finally, U_{atom} represents displacements of individual atoms and ideally includes local displacements only.

U_{atom}



Six parameters

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



One parameter

$$B = 8\pi^2[(U_{11}+U_{22}+U_{33})/3]/10000$$

```
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
```

TLS

research papers

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]

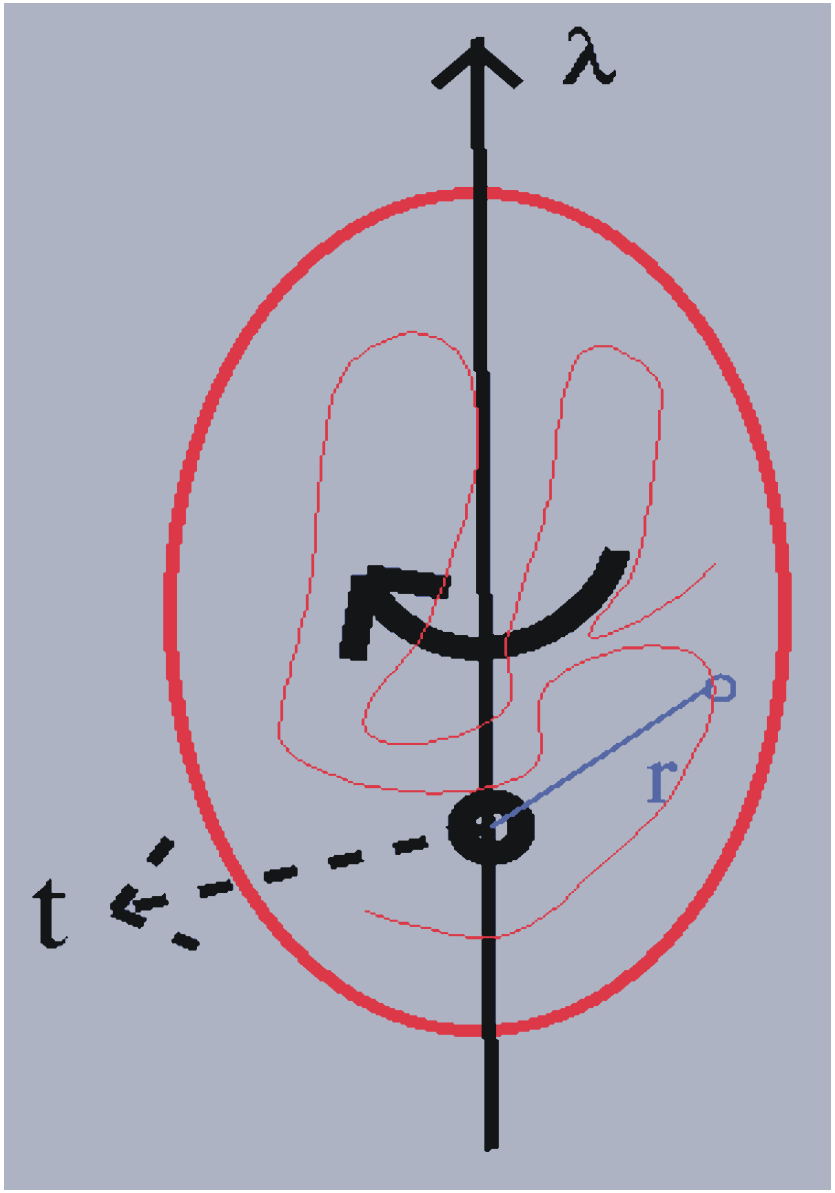
[a]Daresbury Laboratory, Daresbury, Warrington WA4 4AD, England, [b]Department of Chemistry and Biological Sciences, University of Exeter, Exeter EX4 4QD, England, and [c]Chemistry Department, University of York, Heslington, York YO1 5DD, England

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

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

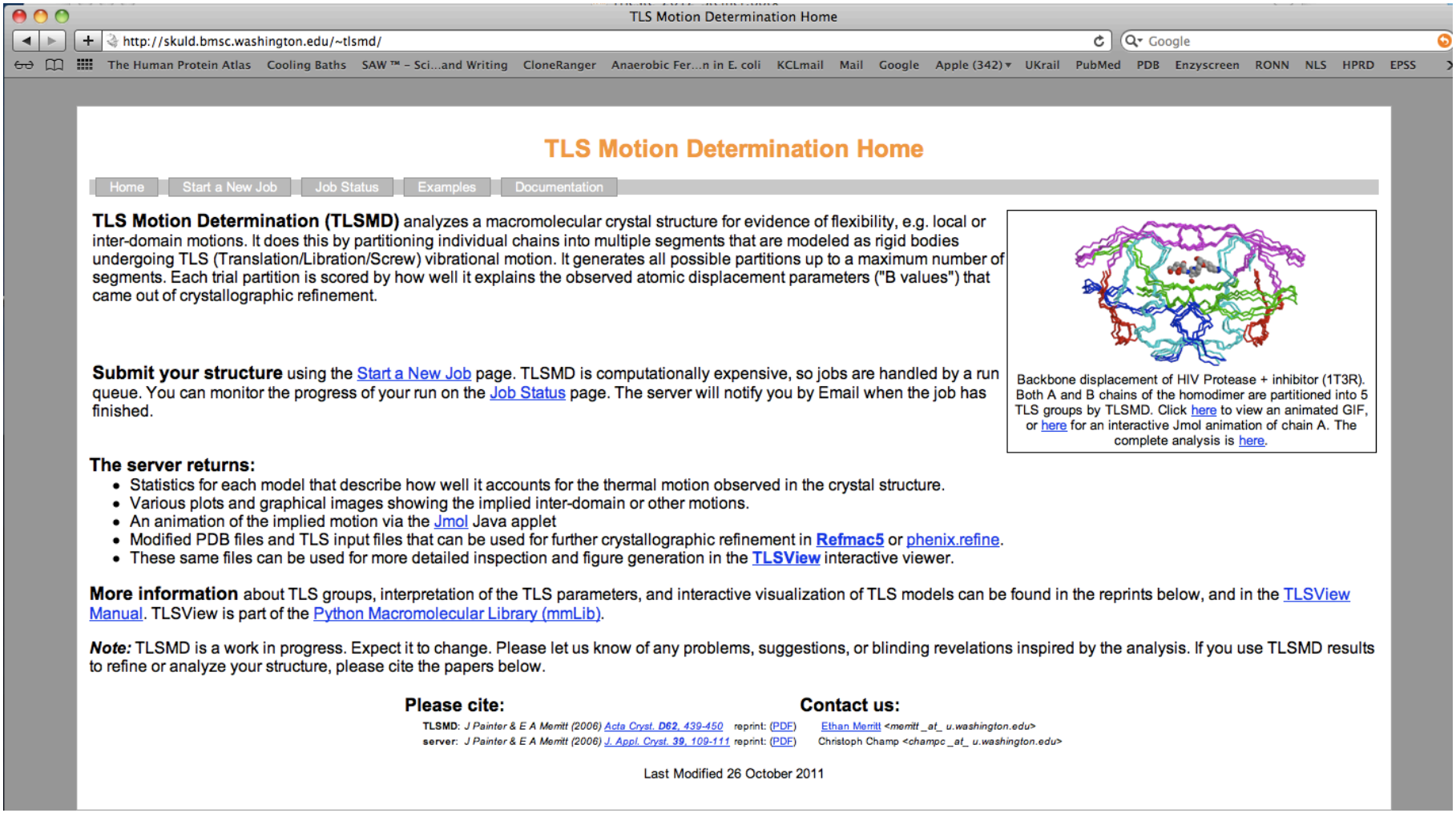
$$U_{TLS} = \langle uu^T \rangle = T + S^T \times \underline{r}^T - \underline{r} \times S - \underline{r} \times L \times \underline{r}^T$$

$T = \langle \underline{tt}^T \rangle$ 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



TLS Motion Determination Home

http://skuld.bmsc.washington.edu/~tlsmd/

The Human Protein Atlas | Cooling Baths | SAW ™ – Sci...and Writing | CloneRanger | Anaerobic Fer...n in E. coli | KCLmail | Mail | Google | Apple (342) | UKrail | PubMed | PDB | Enzyscreen | RONN | NLS | HPRD | EPSS

TLS Motion Determination Home

Home | Start a New Job | Job Status | Examples | Documentation

TLS Motion Determination (TLSMD) analyzes a macromolecular crystal structure for evidence of flexibility, e.g. local or inter-domain motions. It does this by partitioning individual chains into multiple segments that are modeled as rigid bodies undergoing TLS (Translation/Libration/Screw) vibrational motion. It generates all possible partitions up to a maximum number of segments. Each trial partition is scored by how well it explains the observed atomic displacement parameters ("B values") that came out of crystallographic refinement.

Backbone displacement of HIV Protease + inhibitor (1T3R). Both A and B chains of the homodimer are partitioned into 5 TLS groups by TLSMD. Click here to view an animated GIF, or here for an interactive Jmol animation of chain A. The complete analysis is here.

Submit your structure using the Start a New Job page. TLSMD is computationally expensive, so jobs are handled by a run queue. You can monitor the progress of your run on the Job Status page. The server will notify you by Email when the job has finished.

The server returns:

- Statistics for each model that describe how well it accounts for the thermal motion observed in the crystal structure.
- Various plots and graphical images showing the implied inter-domain or other motions.
- An animation of the implied motion via the Jmol Java applet
- Modified PDB files and TLS input files that can be used for further crystallographic refinement in **Refmac5** or phenix.refine.
- These same files can be used for more detailed inspection and figure generation in the **TLSView** interactive viewer.

More information about TLS groups, interpretation of the TLS parameters, and interactive visualization of TLS models can be found in the reprints below, and in the TLSView Manual. TLSView is part of the Python Macromolecular Library (mmLib).

Note: TLSMD is a work in progress. Expect it to change. Please let us know of any problems, suggestions, or blinding revelations inspired by the analysis. If you use TLSMD results to refine or analyze your structure, please cite the papers below.

Please cite:

TLSMD: *J Painter & E A Merritt (2006) Acta Cryst. D62, 439-450* reprint: (PDF)
server: *J Painter & E A Merritt (2006) J. Appl. Cryst. 39, 109-111* reprint: (PDF)

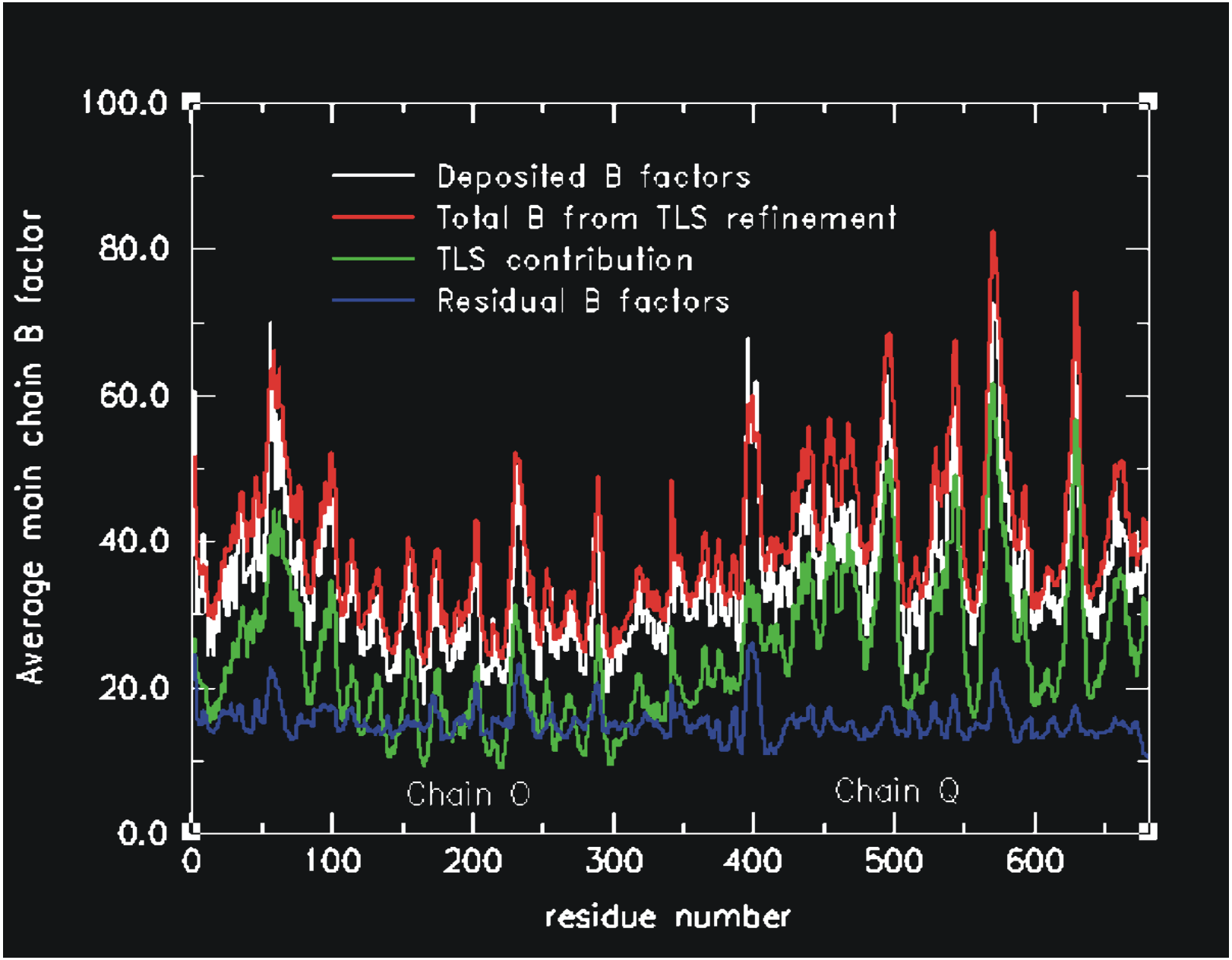
Contact us:

Ethan Merritt *<meritt _at_ u.washington.edu>*
Christoph Champ *<champc _at_ u.washington.edu>*

Last Modified 26 October 2011

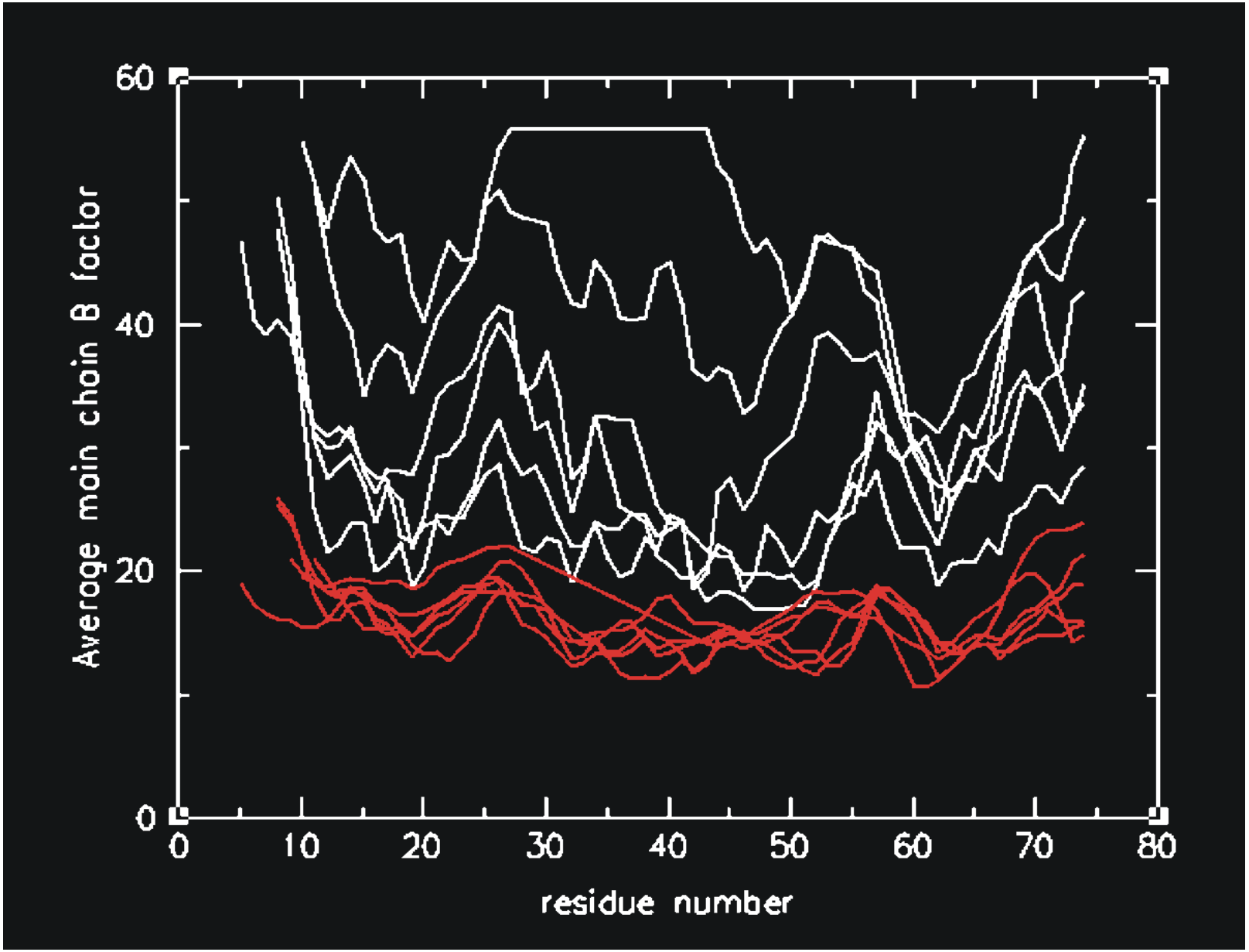
Resolution is not a problem. There are only 20 more parameters per TLS group

Contributions to equivalent isotropic *B*s



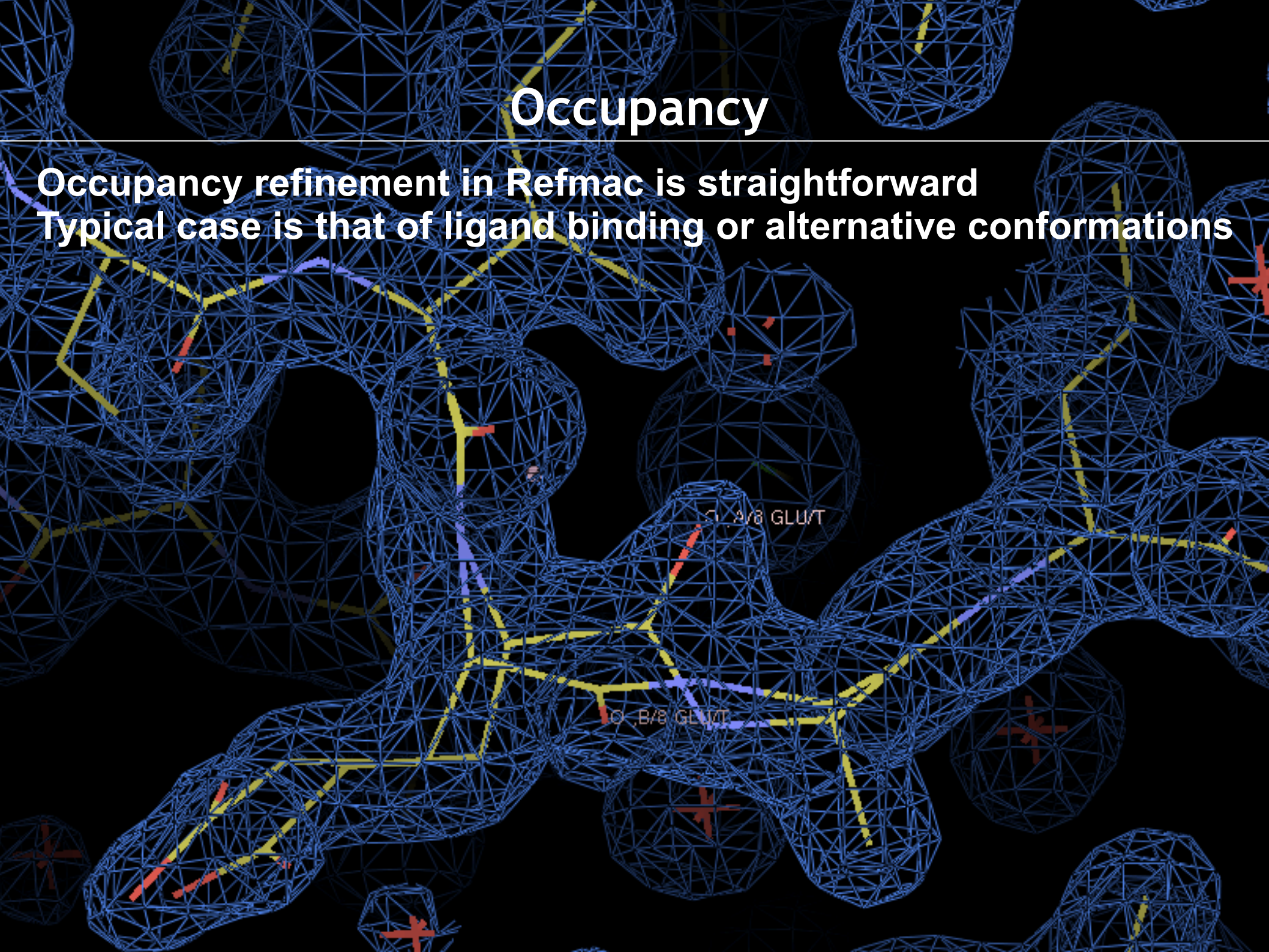
[Howlin, B. & al. (1993) TLSANL: TLS parameter-analysis program for segmented anisotropic refinement of macromolecular structures, J. Appl. Cryst. 26, 622-624]

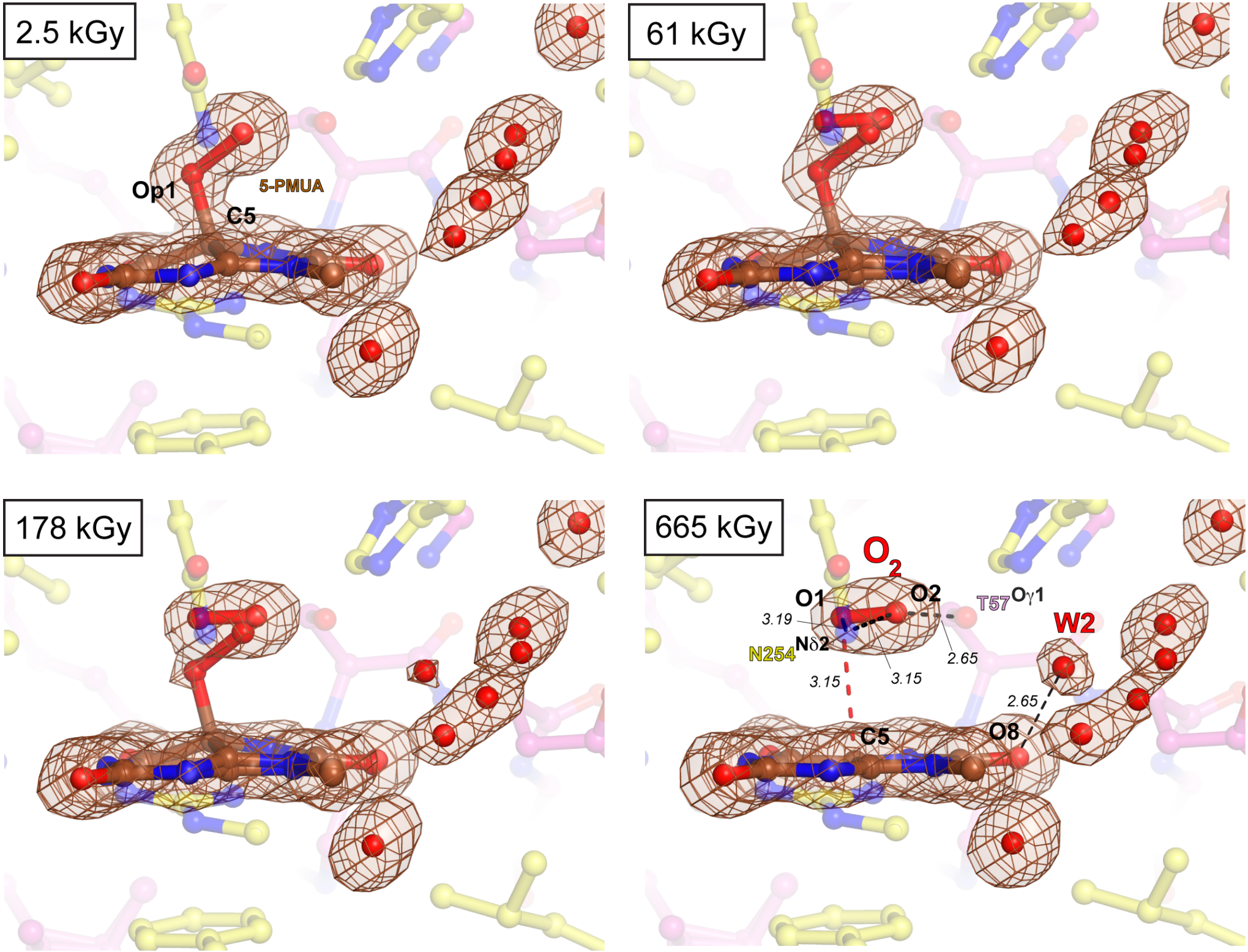
*B*s from NCS related chains



Occupancy

Occupancy refinement in Refmac is straightforward
Typical case is that of ligand binding or alternative conformations




2.5 kGy
Op1
5-PMUA
C5
61 kGy
178 kGy
665 kGy
O2
O1
O2
T57
Oγ1
W2
3.19
Nδ2
N254
2.65
3.15
3.15
2.65
C5
O8

Occupancy

At cryo-temperature alternative conformations reflect static disorder, which in turn is likely a reflection of dynamics in solution.

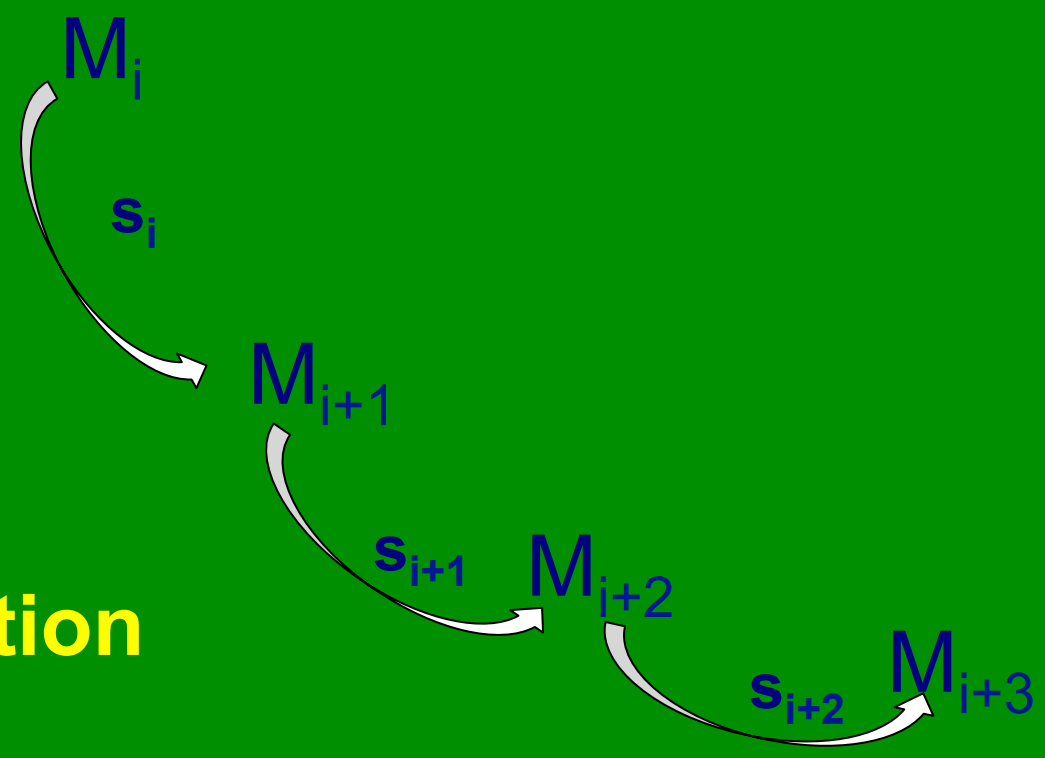
```
ATOM      1  N   AARG A 192      -5.782  17.932  11.414  0.72  8.38      N
ATOM      2  CA  AARG A 192      -6.979  17.425  10.929  0.72 10.12      C
ATOM      3  C   AARG A 192      -6.762  16.088  10.271  0.72  7.90      C
ATOM      7  N   BARG A 192     -11.719  17.007   9.061  0.28  9.89      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
```

In soaking studies partial occupancies are rather common

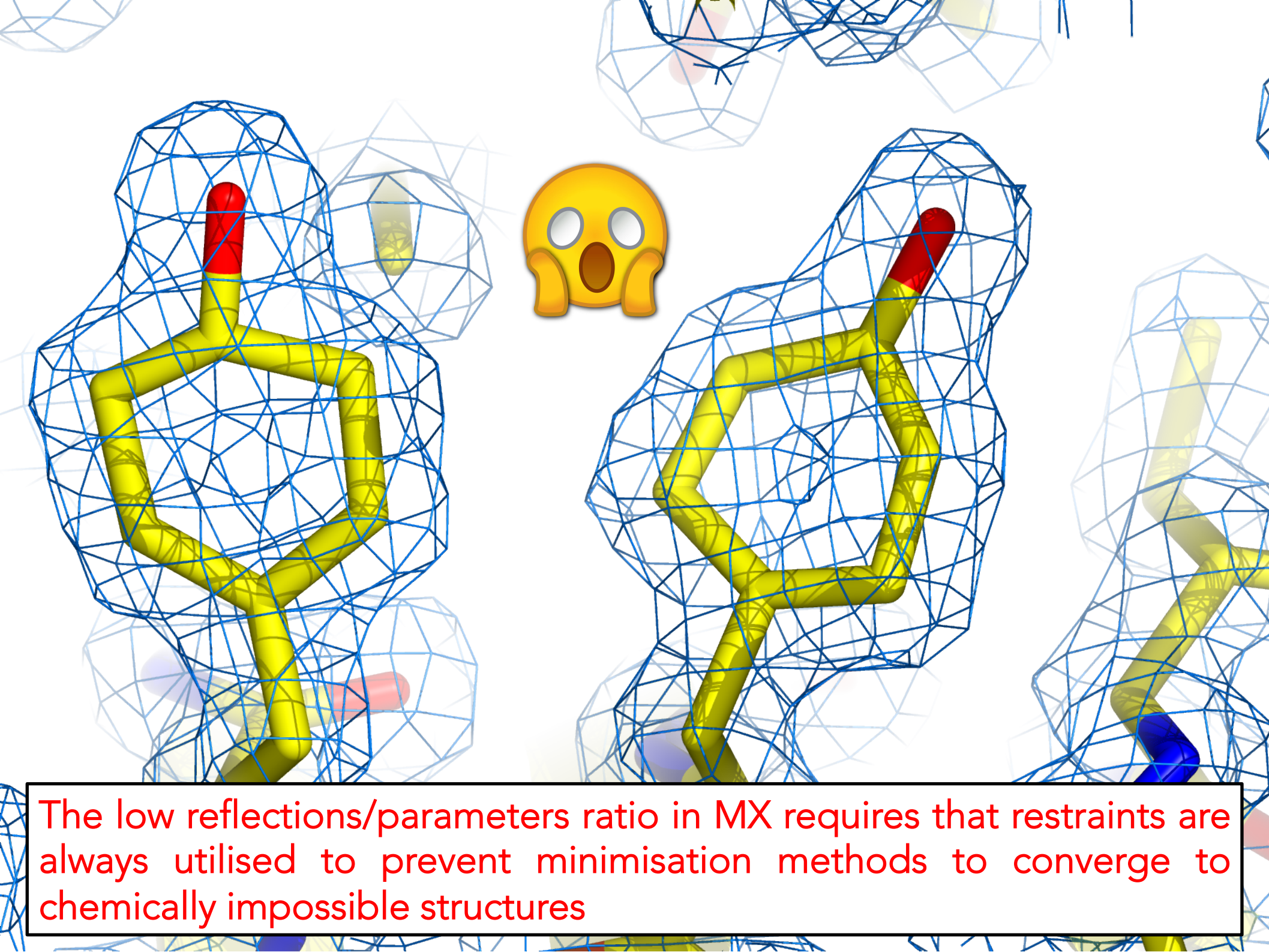
Summary parametrization

- Difficult to give a summary.
- Rigid body/jelly body followed by restrained positional refinement.
- Very low resolution jelly body/DEN.
- 1.4A data or better refine anisotropic ADPs. You should see a significant drop in R values (2-3% or more).
- If you have more than one molecule in a.u. use NCS (local/global).
- If you have a ligand refine its occupancy.

Key aspects of (reciprocal space) refinement

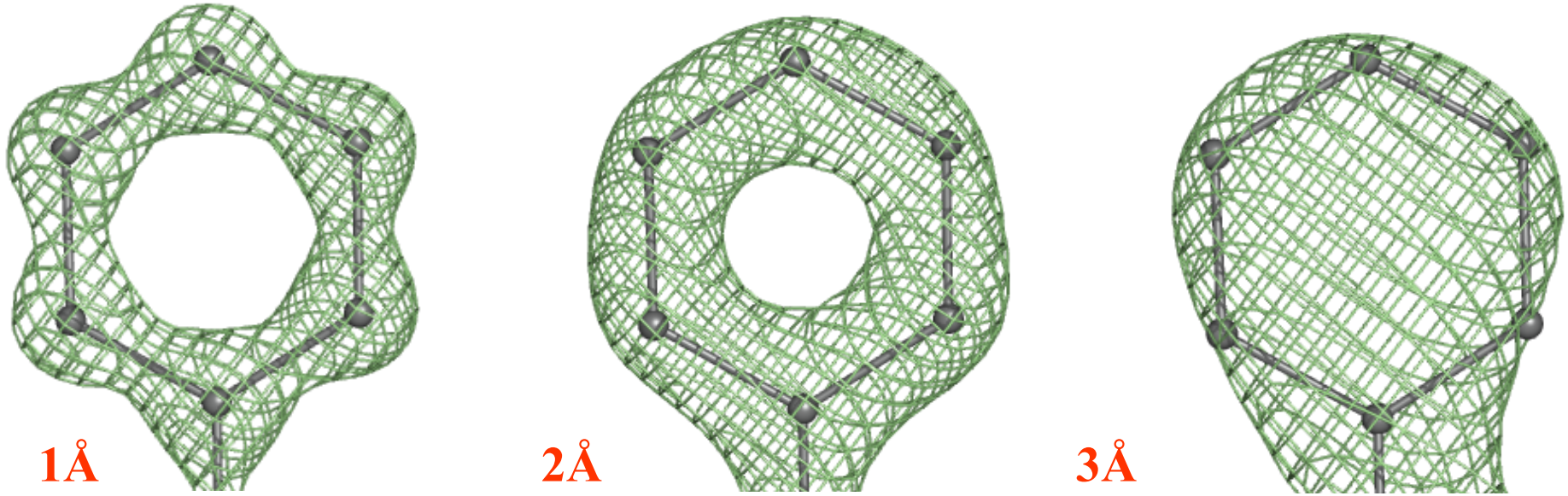


- **Objective function**
- **Method of optimization**
- **Model parametrization**
- **<u>Prior knowledge</u>**



The low reflections/parameters ratio in MX requires that restraints are always utilised to prevent minimisation methods to converge to chemically impossible structures

Restraints



$$f_{\text{total}} = f_{\text{geom}} + w f_{\text{xray}}$$

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

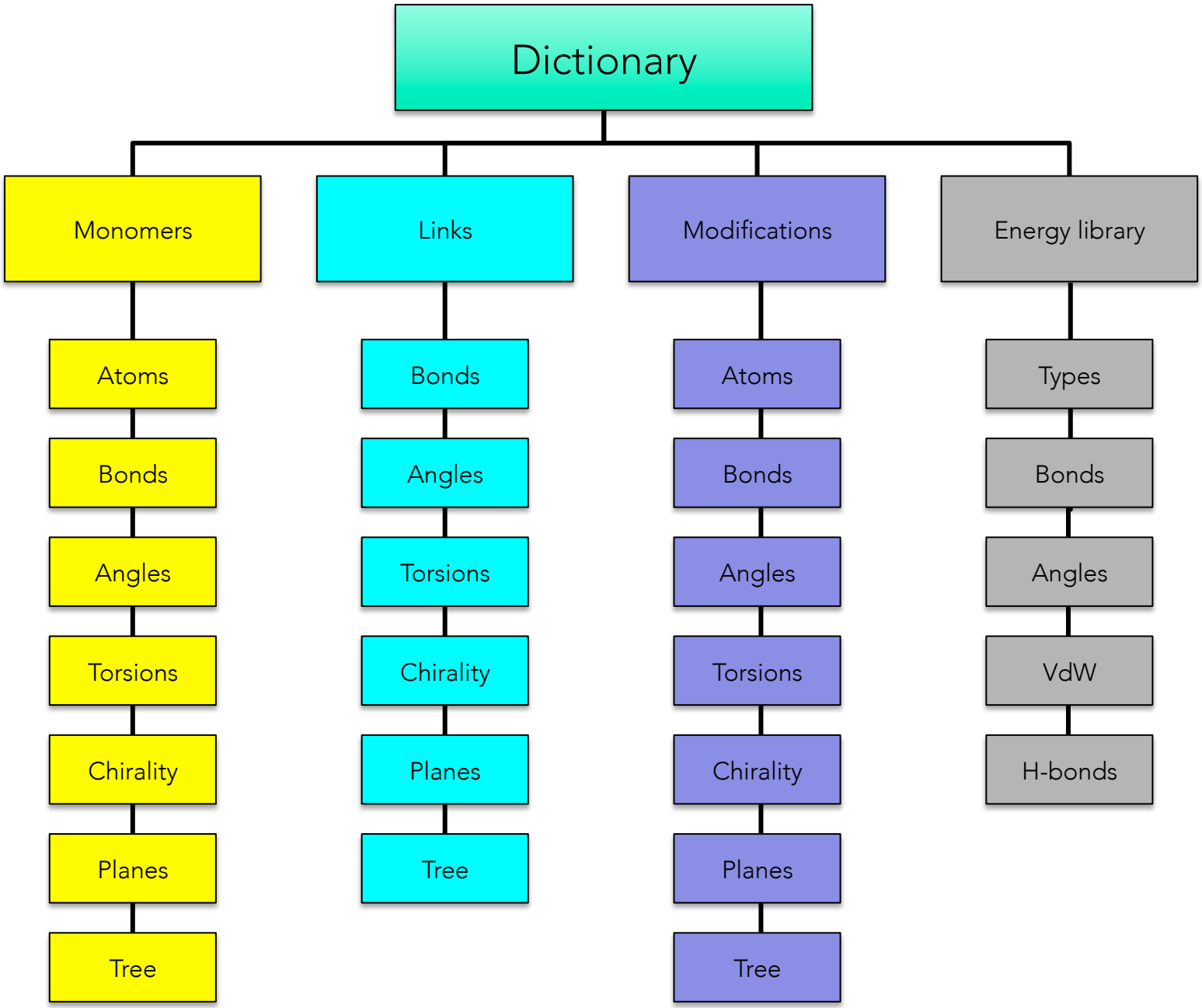
‡ Current address: IFOM – The FIRC Institute of Molecular Oncology, Via Adamello 16, 20139 Milano, Italy

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.

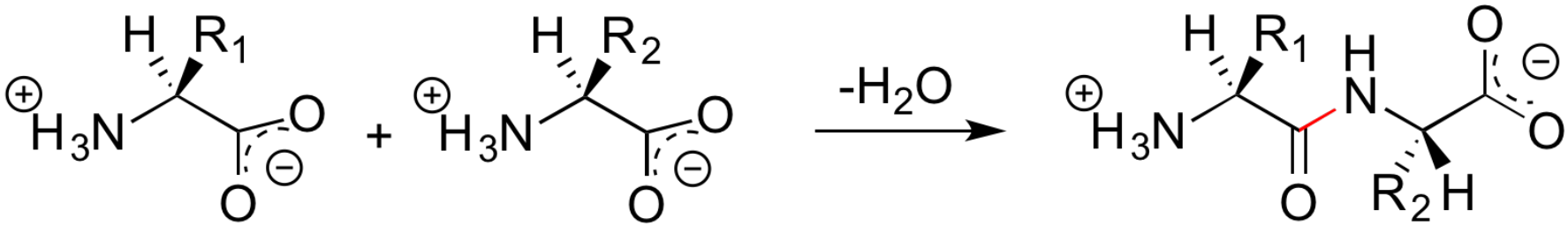
Received 19 April 2004
Accepted 22 September 2004

The use of prior knowledge requires its organised storage.

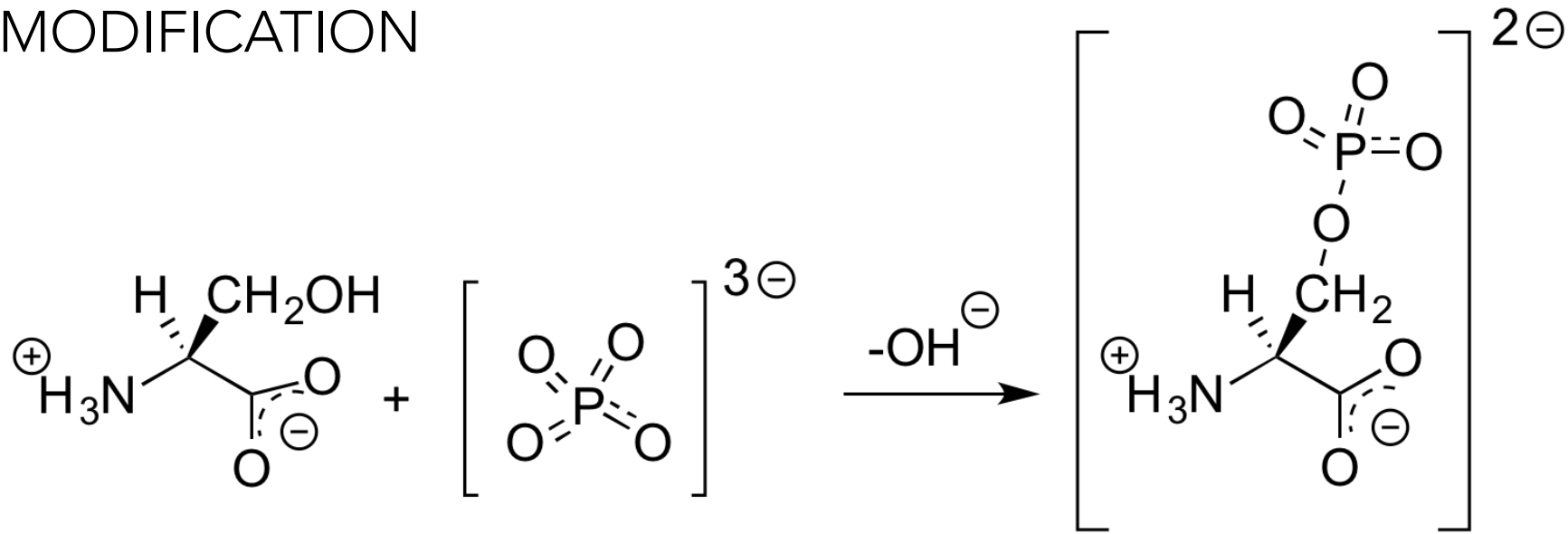

Dictionary
Monomers
Links
Modifications
Energy library
Atoms
Bonds
Angles
Torsions
Chirality
Planes
Tree
Bonds
Angles
Torsions
Chirality
Planes
Tree
Atoms
Bonds
Angles
Torsions
Chirality
Planes
Tree
Types
Bonds
Angles
VdW
H-bonds

Links and Modifications

LINK



MODIFICATION



Current status of the *Refmac*5 dictionary

Used also by *COOT*, *phenix.refine*, *PDB_REDO*

Currently, there are

11617	monomers (complete description)
73	links
63	modifications

These are represented by mmCIF files that can be found in /ccp4-6.5/lib/data/monomers

Description of monomers

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  .
```

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_angle)

```
loop_
_chem_comp_angle.comp_id
_chem_comp_angle.atom_id_1
_chem_comp_angle.atom_id_2
_chem_comp_angle.atom_id_3
_chem_comp_angle.value_angle
_chem_comp_angle.value_angle_esd
 ALA       H       N       CA       114.000     3.000
 ALA       HA      CA      CB       109.000     3.000
 ALA       CB      CA      C        110.500     1.500
 ALA       HA      CA      C        109.000     3.000
 ALA       N       CA      HA       110.000     3.000
 ALA       N       CA      CB       110.400     1.500
 ...
 ...
 ALA       N       CA      C        111.200     2.800
 ALA       CA      C       O        120.800     1.700
```

What happens when you run *REFMAC*5?

If your model only contains monomers for which there is a description

the program takes everything from the library and carries on

You have monomer(s)/link(s)/modification(s) for which there is no description

the program will stop as it needs restraints for the unknown entry/entries

Links / Modifications *JLigand* *(CCP4)*

Ligands

AceDRG (CCP4)
Grade (Global Phasing)
phenix.elbow (*Phenix*)
….

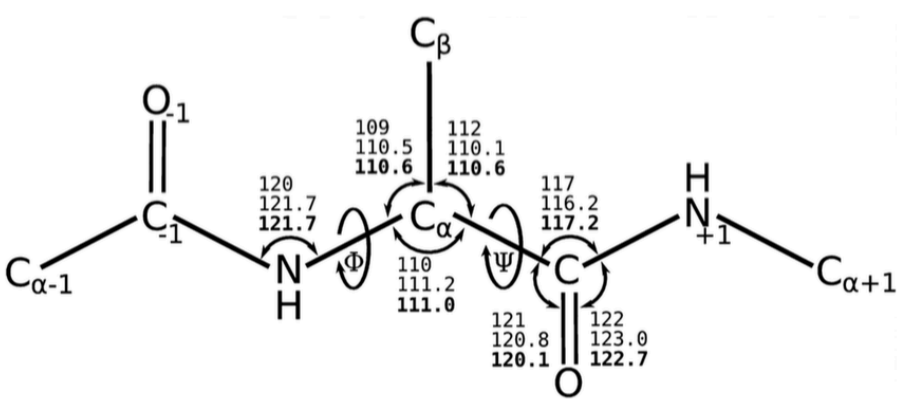
Target restraints

Cambridge Structural Database (CSD) / Crystallography Open Database (COD)
(sub)atomic resolution macromolecules / QM calculations

In the case of proteins:

Engh, R.A., and Huber, R. (1991).
Accurate bond and angle parameters for X-ray protein structure refinement.
Acta Crystallogr. A Found. Crystallogr. 47, 392–400.

Engh, R.A., and Huber, R. (2001).
International Tables for Crystallography. In International Tables for Crystallography,
M.G. Rossmann and E. Arnold, eds. (Dordrecht, The Netherlands: Kluwer Academic Publishers), pp. 382–392.



Single value library (SVL)

taget values are independent of context

Acta Cryst D STRUCTURAL BIOLOGY

ISSN 2059-7983

Keep it together: restraints in crystallographic refinement of macromolecule–ligand complexes

Roberto A. Steiner[a]* and Julie A. Tucker[b]*

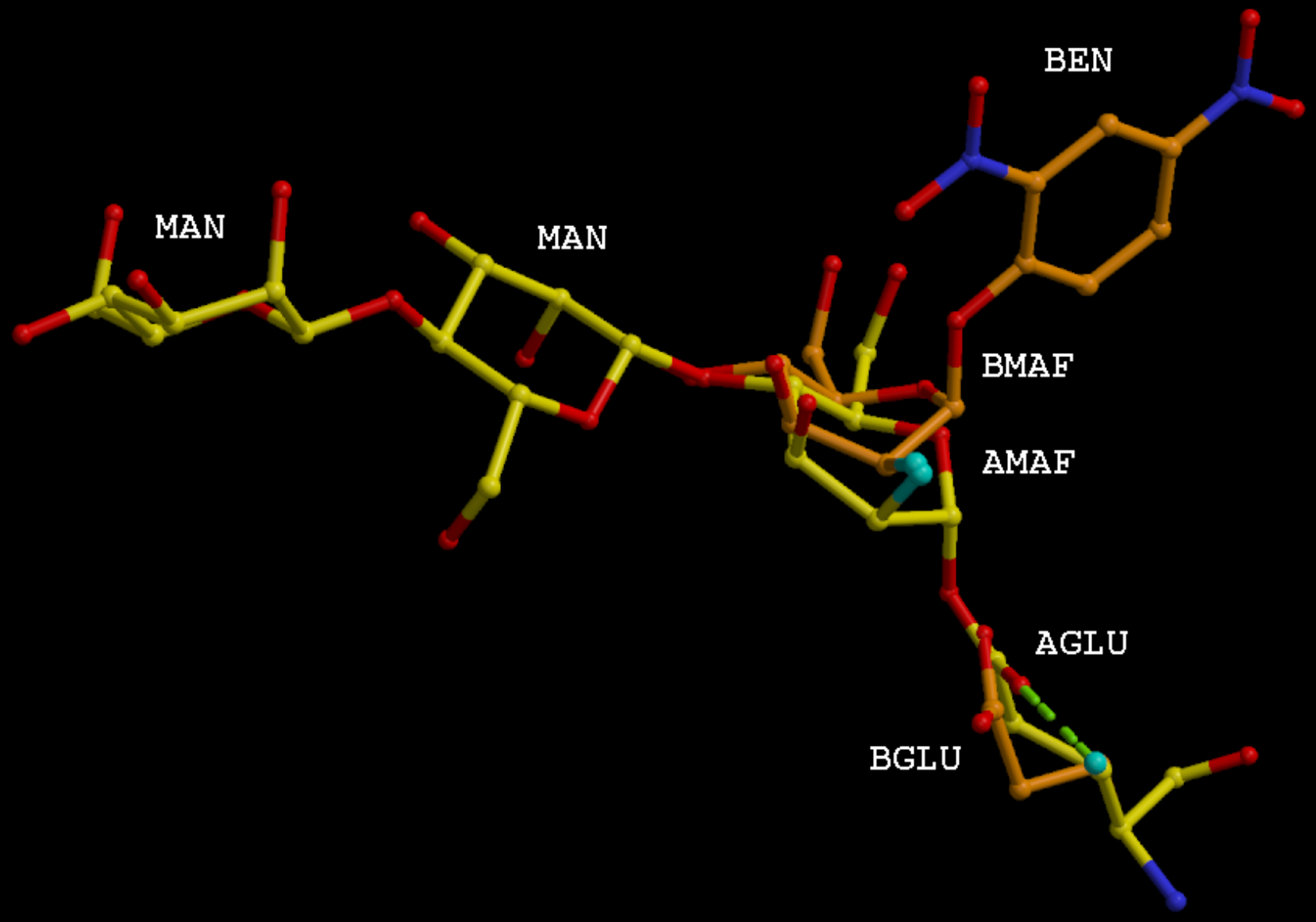
[a]Randall Division of Cell and Molecular Biophysics, King's College London, London SE1 1UL, England, and [b]Northern Institute for Cancer Research, Paul O'Gorman Building, Medical School, Newcastle University, Framlington Place, Newcastle-upon-Tyne NE2 4HH, England. *Correspondence e-mail: julie.tucker@newcastle.ac.uk, roberto.steiner@kcl.ac.uk

Received 30 September 2016
Accepted 8 November 2016

Keywords: restraint sets; ligand complexes; standard deviation; macromolecular crystallography; refinement.

A short introduction is provided to the concept of restraints in macromolecular crystallographic refinement. A typical ligand restraint-generation process is then described, covering types of input, the methodology and the mechanics behind the software in general terms, how this has evolved over recent years and what to look for in the output. Finally, the currently available restraint-generation software is compared, concluding with some thoughts for the future.

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
12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789
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 *CCP*4 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

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 *CCP*4 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 *CCP*4 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.

Received 22 November 2011
Accepted 19 January 2012

Few final remarks

- Your original reflection file should always be your MTZIN
- The MTZOUT is used only for map calculations
- If you have phase information (HL coefficents) use it at the early/ medium stage of refinement then drop it. Same goes for SAD/SIRAS data.
- I tend to include hydrogens (riding) at let's say resolution better than 3A. Do this once the model is quite complete.
- TLS. Generally quite useful. Sometimes you get stunning stats. I use TLSMD to get TLS groups.
- Since the introduction of NCS local I rarely had to employ NCS global.
- Ligands. Often source of problems. Read Steiner and Tucker.
- JLigand (Andrey Lebedev) is extremely convenient to define links.
- Low resolution tools quite powerful (map sharpening, jelly body)
- A fast program makes everything a lot more convenient. PDB_REDO, ARP/wARP,...