# Macromolecular refinement with SHELXL-97

Göttingen, December 9th 2010

George M. Sheldrick

http://shelx.uni-ac.gwdg.de/SHELX/

# R-factors

**Contrary to popular opinion, the aim of a crystallographic refinement is not to reduce the *R*-factor:**

$$R = \Sigma \, |F_o - F_c| \, / \, \Sigma \, F_o$$

**A better indication of the quality of the agreement between the model and the data is obtained by reserving say 5% of the reflections for calculating the *free R-factor* (Brunger, 1992), and not using them actively in the refinement. The free *R* appears to correlate well with the mean phase error.**

**The free *R* is a good global indicator, but small changes in the model often change *it* by less than its esd, which can be large (0.1-0.5%) because it is based on a small number of data. For small molecules there are usually not enough data to estimate a statistically significant free *R*, but the high data to parameter ratio means that overfitting is less of a problem.**

# The ratio $R_{free}$ / $R_{work}$

$R_{free}$ is always larger than the *R*-factor $R_{work}$ based on the working set. A large difference between these two *R*-factors is normally taken as evidence that the structure has been ***over-refined***. This can be the case when going anisotropic or adding waters reduced $R_{work}$ much more than $R_{free}$. The ratio of $R_{free}$ to $R_{work}$ also depends strongly on the resolution. Tickle (2000) showed that to a reasonable approximation:

$$R_{free} \,/\, R_{work} = \{ ( N + P ) / ( N - P ) \}^{½}$$

where ***N*** is the number of reflections and ***P*** the ***effective*** number of parameters refined, allowing for restraints. However both ***N*** and ***P*** are proportional to the number of atoms provided that the resolution ***d*** and solvent content ***s*** are taken into account, leading to the approximation:

$$R_{free} \,/\, R_{work} = \{ ( 1 + Q ) / ( 1 - Q ) \}^{½} \quad \text{where} \quad Q = 0.025\, p\, d^{3} ( 1 - s )$$

***p*** is the effective number of parameters per atom (say 1.5 for isotropic).

# The ratio $R_{free}$ / $R_{work}$ (continued)

$$R_{free} / R_{work} = \{ ( 1 + Q ) / ( 1 - Q ) \}^{½} \quad \text{where} \quad Q = 0.025\, p\, d^3 ( 1 - s )$$

**This expression is evaluated as a function of the resolution *d* for three typical cases (restrained protein and unrestrained small-molecule):**

| | *s* = 0.5, *p* = 1.5<br>isotropic protein | *s* = 0.3, *p* = 4.0<br>anisotropic protein | *s* = 0.1, *p* = 9.0<br>small molecule |
|---|---|---|---|
| *d* = 0.50 | 1.002 | 1.009 | 1.026 |
| *d* = 0.75 | 1.008 | 1.030 | 1.089 |
| *d* = 1.00 | 1.019 | 1.073 | 1.228 |
| *d* = 1.25 | 1.037 | 1.148 | 1.519 |
| *d* = 1.50 | 1.065 | 1.272 | 2.306 |
| *d* = 1.75 | 1.106 | 1.484 | |
| *d* = 2.00 | 1.163 | 1.883 | |
| *d* = 2.25 | 1.242 | 2.978 | |
| *d* = 2.50 | 1.352 | | |
| *d* = 2.75 | 1.509 | | |
| *d* = 3.00 | 1.747 | | |

# Why are R-factors so high (for macromolecules)?

**Typical small molecule *R*-factors are in the range 3-5% and values below 2% are sometimes obtained. However, even when the same data collection hardware and software are employed, the values for proteins are at least 5 times as high.**

**After merging equivalents it is possible to obtain data to a precision of about 1%, as required for successful S-SAD experiments.**

**So why are the *R*-factors for macromolecules so high?**

# The solvent model

**The solvent accounts for 50% of the diffracted intensity (at low resolution) and probably makes a large contribution to the high *R*-factors, but is often modelled with only two parameters. Despite many attempts, we still do not have a really adequate solvent model.**

**Far from the macromolecule, it would be reasonable to assume that the solvent is 'flat'. Closer in, we need to worry about multiple conformations of the molecular periphery and different electrostatics and hydrophilicities, and the fact that the solvent (or soup) is by no means pure water, and may contain disordered PEGs, ions etc.**

**For example, chloride ions would tend to concentrate close to positively charged lysine or arginine side-chains and have a higher mean electron density (ca. 0.47 e/Å$^3$) than water (0.33 e/Å$^3$).**

# Using prior knowledge

***Prior knowledge*** **is essential for successful refinement at poor data to parameter ratios, i.e. low resolution. It is usually used in the form of *restraints*, which can be treated like additional experimental data in both the least-squares and maximum-likelihood approaches.**

**Typical restraints involve bond lengths, angles, planarity etc. These are *unimodal* because they have a single target value.**

**It is important to keep some prior information in reserve for validation purposes. It is convenient to use *multimodal* information such as torsion angles (Ramachandran diagram) for validation because it is less suitable for refinement.**

# Constraints and restraints

**Constraints are exact mathematical conditions that lead to a reduction in the number of parameters. Examples are rigid groups and riding hydrogen atoms.**

**Restraints are additional observational equations involving target values *T* and their standard deviations σ that are added to the quantity to be minimised:**

$$M = \sum w_x(F_o^2 - F_c^2)^2 + \sum w_r(T_{target} - T_c)^2$$

**To bring the X-ray weights $w_x$ onto an absolute scale, in SHELXL they are normalized so that the mean $w_x(F_o^2-F_c^2)^2$ is unity. $w_r = 1/\sigma^2$ should then be structure and resolution independent. A useful side-effect is that $w_x$ increases as the agreement between $F_o^2$ and $F_c^2$ improves during the course of refinement.**

**In REFMAC the user has to set the relative weights of the intensity data and restraints.**

# Types of constraint in SHELXL-97

**Constraints for special positions: the necessary constraints on coordinates, occupancies and $U_{ij}$ are derived automatically.**

**Rigid groups (AFIX 6 … AFIX 0): the 3 positional parameters per atom are replaced by 3 rotations and 3 translations for the whole rigid group. Atoms may not be in more than one rigid group.**

**Riding hydrogen atoms (AFIX mn): $x_H = x_C + \Delta x$**
**– no extra positional parameters.**

**Fixed parameters: just add 10 to *x*, *y*, *z*, *occ*, *U* etc. Typically occupancies are fixed at 1.0 by adding 10, i.e. given as 11.0**

**Free variables: can be used to add extra linear constraints to the usual refinement parameters and also be used instead of restraint target values, e.g. the $C_\alpha$ chiral volumes of all proline residues could be restrained to be equal to the same free variable. This provides a convenient way of getting target values with esds for use as restraints in other structures.**

# Types of restraint in SHELXL-97

**DFIX, DANG and SADI - distances and ‘angle distances’**

**FLAT and CHIV - planarity and chiral volumes**

**BUMP - antibumping**

**NCSY - non-crystallographic symmetry (NCS)**

**DELU, SIMU and ISOR - (an)isotropic displacements**

**SUMP - general ‘free variable’ restraint (e.g. for the sum of occupancies of side-chains with three disorder components)**

**DEFS sets default restraint esds and SAME can generate SADI restraints. CHIV, BUMP, SAME, NCSY and DELU make use of the connectivity array.**

# Least-squares algebra and standard uncertainties

In non-linear least-squares refinement, the parameter shifts each cycle are calculated by $\delta x = A \cdot B^{-1}$ where:

$$A_j = \sum w(F_o^2 - F_c^2)(\partial F_c^2 / \partial x_j) \quad \text{and} \quad B_{ij} = \sum w(\partial F_c^2 / \partial x_i)(\partial F_c^2 / \partial x_j)$$

where the summations are over all reflections. The esds (now called ***standard uncertainties*** by the IUCr) are then given by:

$$\text{esd}(x_j) = [\ (B^{-1})_{jj} \sum w(F_o^2 - F_c^2)^2 / (N_R - N_P)\ ]^{1/2}$$

provided that $\sum w(F_o^2 - F_c^2)^2$ is normally distributed, i.e. shows no systematic trends with intensity, resolution and other factors. $N_R$ is the number of reflections and $N_P$ the number of parameters.

These estimates only take random errors into account. Since systematic errors can never be completely eliminated, such esds are always underestimated. Comparison of independent determinations of the same small molecule structures suggest that coordinate esds are underestimated by a factor of about 1.5 and esds in $U$ or $U_{ij}$ by a factor of about 2. Nevertheless, for a single structure determination they are by far the best estimates that we have.

# How to estimate esds for macromolecules

**Collect data to as high a resolution as possible (say < 1.5Å).**

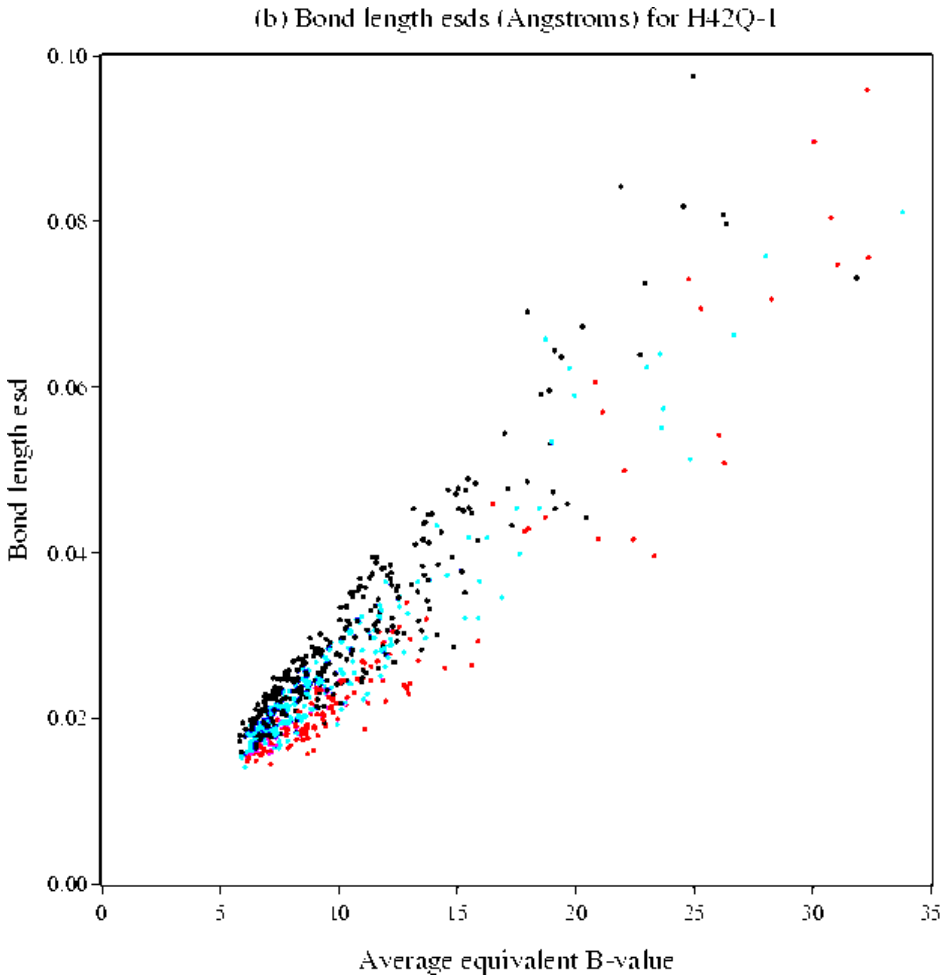
**Refine to convergence with CGLS, first using only the working set and finally with all data (without adding extra parameters).**

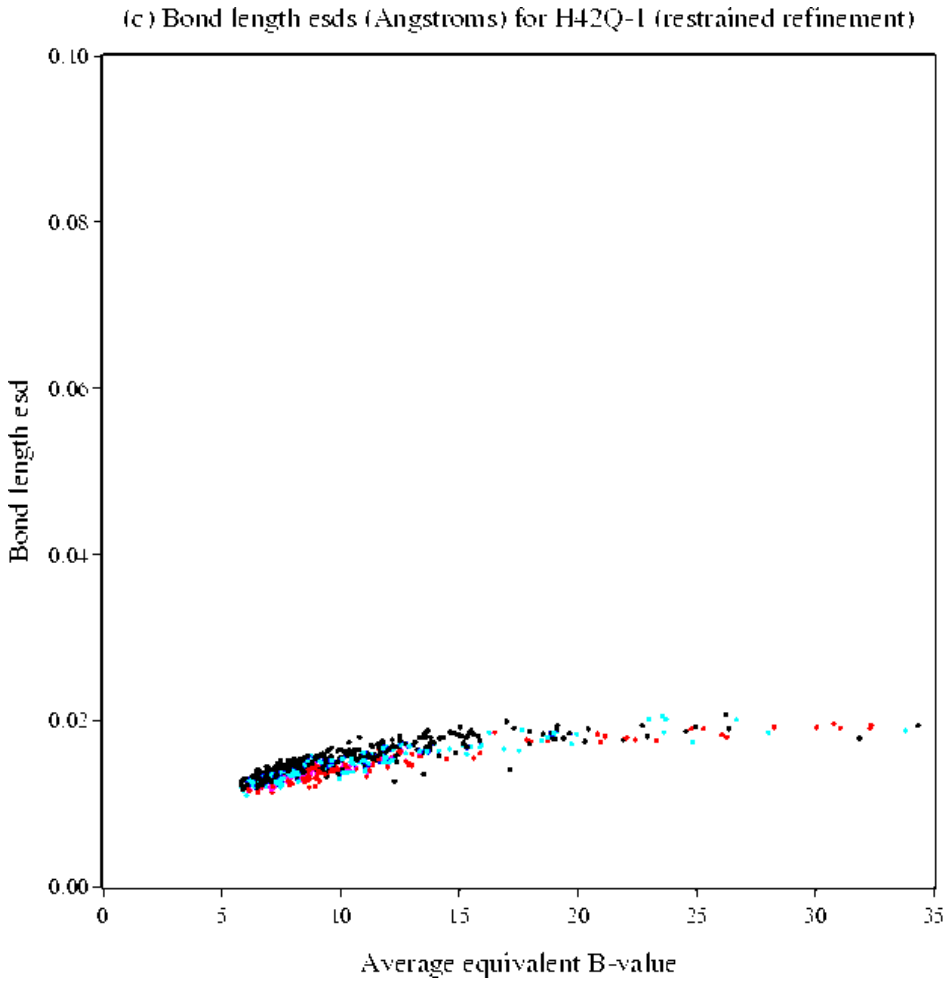
**Perform one final full-matrix cycle with zero damping and zero shift multiplier (L.S. 1 and DAMP 0 0). Switch off all restraints. Restraints and Marquardt damping would lead to under-estimated esds.**

**If the full-matrix refinement would require the purchase of extra memory, an adequate compromise is BLOC 1 to define a full-matrix block consisting of all geometrical but no displacement parameters.**

**SHELXL uses the full covariance matrix and the estimated unit-cell errors to estimate the *standard uncertainties* (*esds*) in all dependent parameters.**

# Esds in bond lengths (unrestrained and restrained)




**The unrestrained bond length esds (left, color coded C-C black, C-N blue and C-O red) show the expected dependence on atomic number and equivalent *B*; the restrained esds (right) rise asymptotically to 0.02Å, the esd used on the DFIX instructions.**

# How to weight the restraints

**The bond length and angle restraints of Engh & Huber (2001), derived from the CSD, are almost universally employed for proteins. With SHELXL and REFMAC it is usual to set the restraint esds to about 0.02 Å for all distances. The resulting rms deviations from the ideal bond lengths after refinement are about 0.0090 Å for SHELXL and 0.0165 Å for REFMAC, according to a PDB survey by Jaskolski et al. (2007), who also suggested that the restraints should be relaxed for the well defined parts of high-resolution protein structures. However their paper was criticized by Stec (2007) and Tickle (2007).**

**There would be a better case for reducing the restraint esds for bond lengths that were more tightly distributed and loosening them for the distances that showed a larger spread in Engh and Huber's study, as indeed intended by them.**


**Engh R.A. & Huber, R. (2001). *Int. Tables for Crystallography,* vol. F, pp.382-392.**

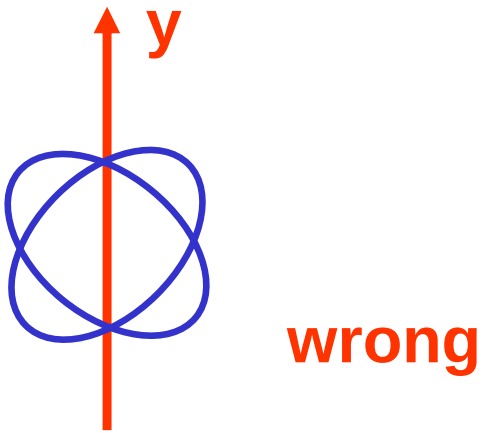
**Jaskolski, M., Gilski, M., Dauter, Z. & Wlodawer, A. (2007). *Acta Cryst*. D63, 611-620 and 1282-1283.**

**Stec, B. (2007). *Acta Cryst*. D63, 1113-1114.**

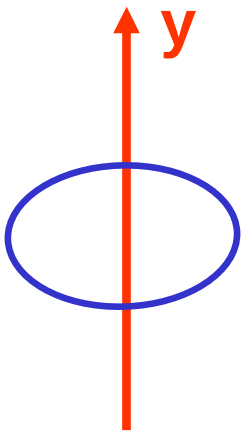
**Tickle, I. (2007). *Acta Cryst*. D63, 1274-1281.**

# Special position constraints

**Example:** **Atom on twofold axis in space group C2. The two positions related by the twofold axis (*x*,*y*,*z*: -*x*,*y*,-*z*) coincide when *x* = 0 and *z* = 0. Since we still wish to sum over all symmetry operators in the structure factor calculation, the occupancy is fixed at 0.5. The probability ellipsoid used to describe the anisotropic motion should not be changed by the 180º rotation:**




**wrong**




**right**

**[$U_{11}$, $U_{22}$, $U_{33}$, $U_{23}$, $U_{13}$, $U_{12}$] ≡ [$U_{11}$, $U_{22}$, $U_{33}$, -$U_{23}$, $U_{13}$, -$U_{12}$], which is only true if $U_{23}$ = 0 and $U_{12}$ = 0.**

**All these *constraints* are generated automatically by SHELXL for all special positions in all space groups.**

# Rigid group constraints

**In SHELXL, rigid groups are defined by three rotations about the first atom in the group and by three translations of the group as a whole. Special position constraints may be applied to the first atom and restraints and riding hydrogens are allowed on all atoms in the group. Full matrix refinement is essential for rigid groups because of the strong parameter correlations involved, but the number of parameters involved is small, e.g. in the first refinement step after a MR solution of a protein. Note that the esds of bond lengths and angles but not of coordinates within a rigid group come out as zero from the L.S. matrix algebra.**

**AFIX 6** rigid group – all
**…** bond lengths and
**atoms** angles fixed
**…**
**AFIX 0**

**AFIX 9** variable metric
**…** rigid group – angles fixed,
**atoms** bond lengths multiplied by
**…** the same factor
**AFIX 0**

# Free variables

**Free variables are an extremely concise but effective way of applying linear constraints to atom parameters (especially occupancies), restraint targets etc. The parameter *x* is given as (10*m*+*p*), which is interpreted as follows:**

***m* = 0: refine normally, starting at value *p***

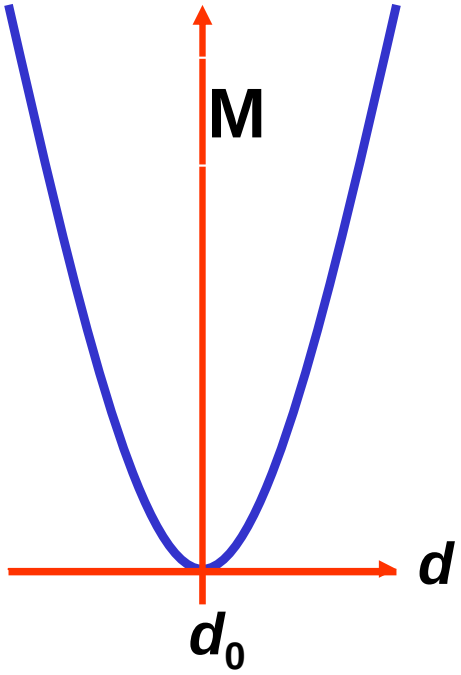
***m* = 1: fix at value *p***

***m* > 1: *x* = *p**fv(*m*)**

***m* <–1: *x* = *p**[fv(–*m*) – 1]**

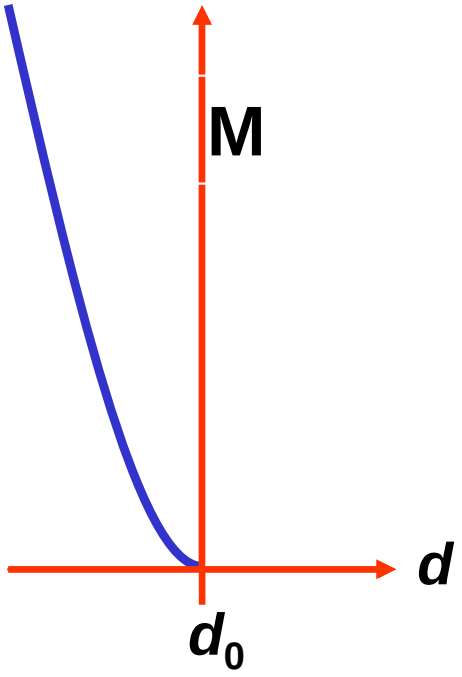
**e.g. 30.25 (m = 3, p = 0.25) means 0.25*[fv(3)] and –30.25 (m = –3, p = –0.25) means 0.25*[1 – fv(3)], which could be used to constrain two occupancies to add up to 0.25 (only one parameter, free variable #3, is refined). The starting values for the free variables are given on the FVAR instruction (but free variable #1 is the overall scale factor).**

# Distance and antibumping restraints




DFIX
DANG
} + $\Sigma\ (1/\sigma^2)\ (d - d_0)^2$

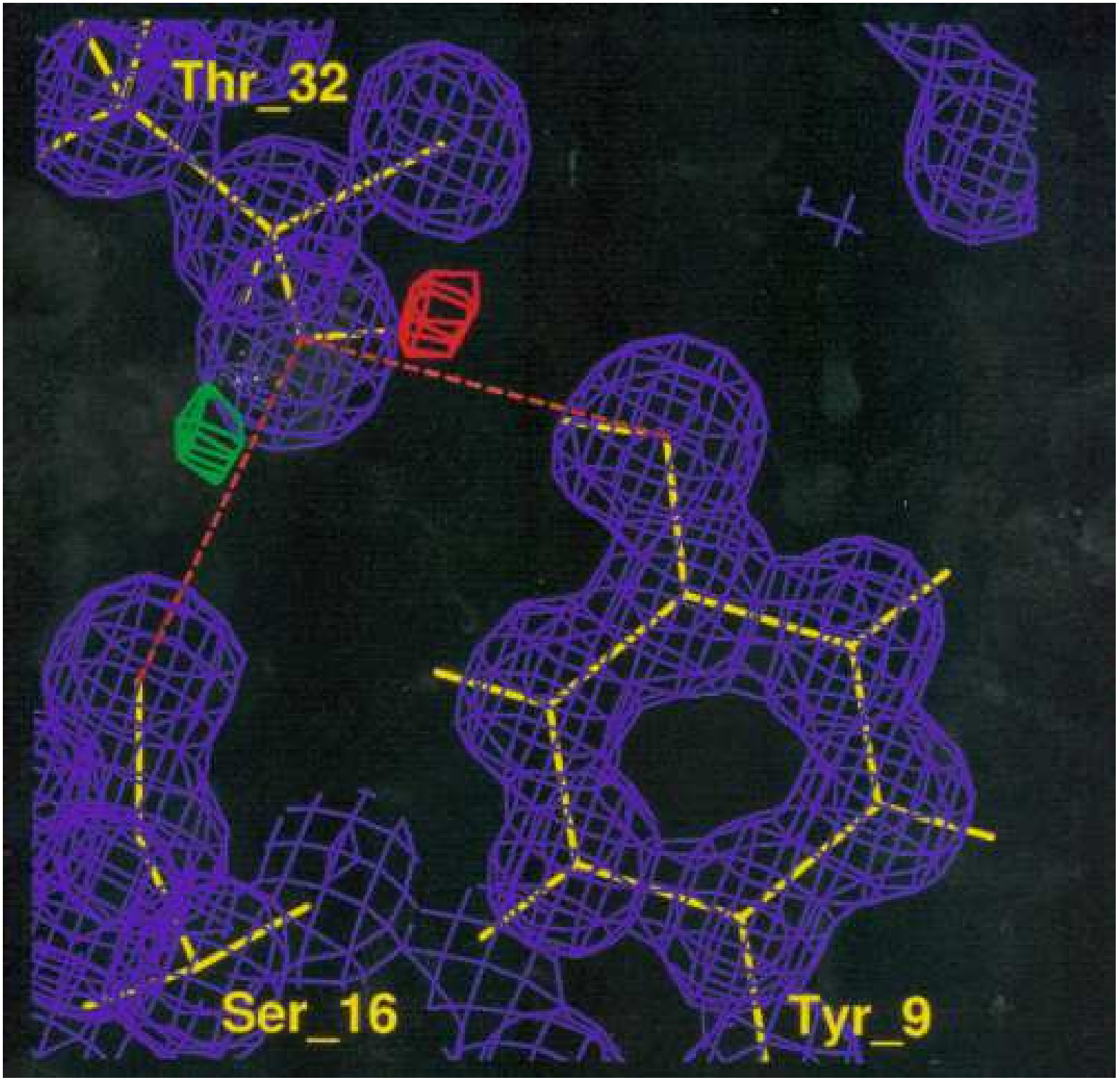
SADI + $\Sigma\ (1/\sigma^2)\ (d_1 - d_2)^2$




**BUMP**

**+ $\Sigma\ (1/\sigma^2)\ (d - d_0)^2$**

**if $d < d_0$**




# A wrongly placed H-atom

Green: +2.3σ
Red: -2.3σ

Thomas R. Schneider

# Similar distance restraints

**Similar distance restraints assume that distances are equal, but without target values (which might turn out to be inaccurate). E.g. a structure contains 6 phosphate anions and we wish to refine them as regular tetrahedra with equal bond lengths, but we don't know what target value to use (it will be affected by pH and by libration):**

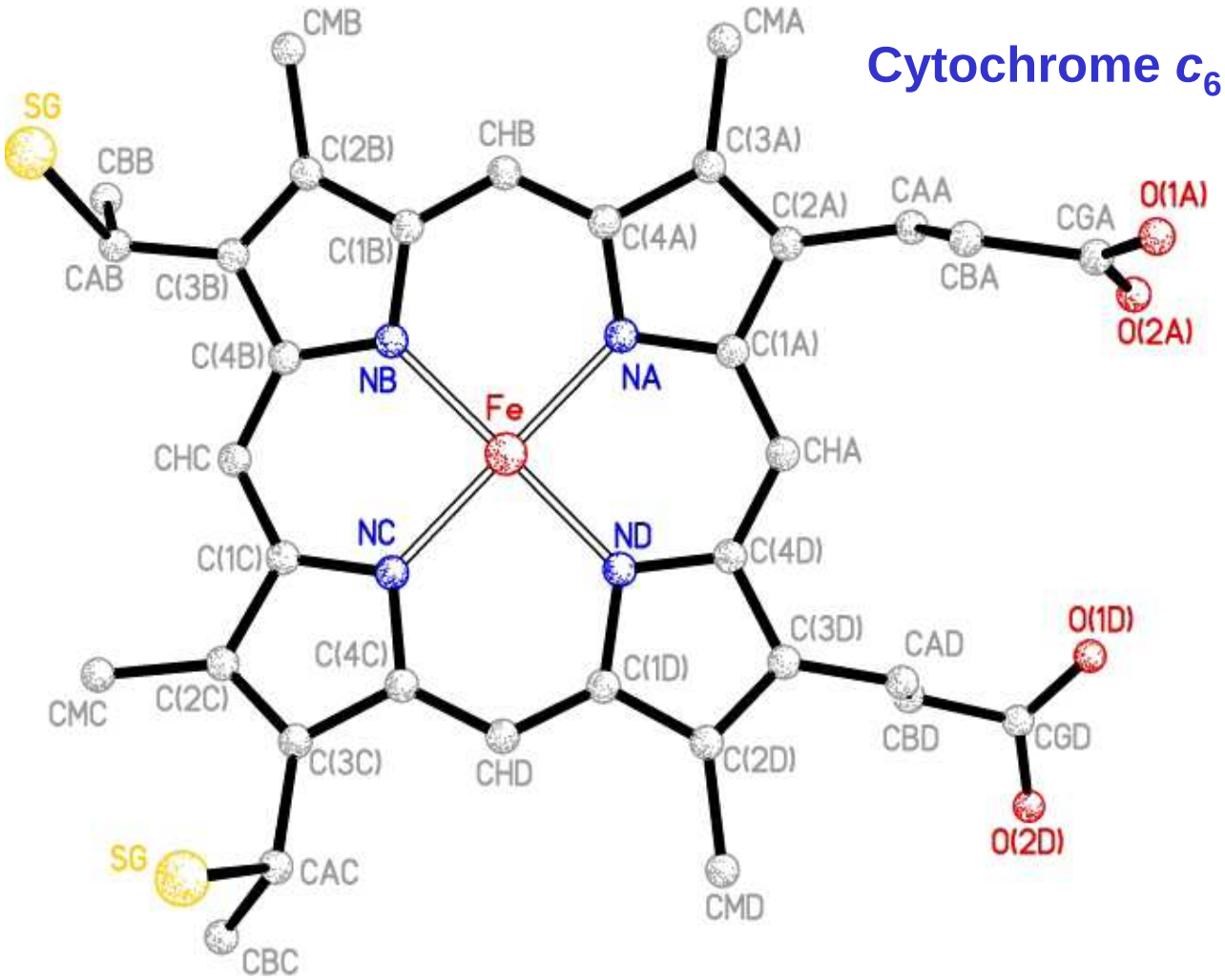
```
SADI_PO4  P O1  P O2  P O3  P O4

SADI_PO4  O1 O2  O1 O3  O1 O4  O2 O3  O2 O4  O3 O4
```

**The first line restrains all 6x4 P–O distances to be equal, the second equates the 6x6 O•••O distances.**

**SADI instructions may also be generated automatically by SAME, but this is accident-prone so not recommended.**

# Use of free variables to obtain mean distances with esds




The following input refines fv 2, 3 and 4 to be the mean Fe-N, N-C and N…CH distances. Because of the 4- and 8-fold redundancy, accurate values are obtained that can be used as restraints.

```
FVAR 1.0 1.8 1.4 2.4
DFIX_HEM 21 Fe NA Fe NB Fe NC Fe ND
DFIX_HEM 31 NA C1A NA C4A NB C1B NB C4B NC C1C NC C4C ND C1D ND C4D
DFIX_HEM 41 NA CHA NA CHB NB CHB NB CHC NC CHC NC CHD ND CHD ND CHA
etc...
```

# Comparison of target bond lengths

| | Traditional | Cytochrome $c_6$ | CSD |
|---|---|---|---|
| NA–C1A | 1.384 | 1.382(20) | 1.381(18) |
| C1A–CHA | 1.378 | 1.376(17) | 1.379(17) |
| C1A–C2A | 1.449 | 1.445(23) | 1.443(15) |
| C2A–C3A | 1.334 | 1.347(23) | 1.355(19) |
| NA•••C2A | 2.312 | 2.314(14) | 2.318(14) |
| NA•••CHA | 2.450 | 2.438(15) | 2.442(16) |
| C1A•••C4D | 2.456 | 2.456(7) | 2.462(17) |
| C1A•••C4A | 2.211 | 2.203(14) | 2.197(20) |
| C1A•••C3A | 2.247 | 2.248(7) | 2.248(16) |
| C2A•••CHA | 2.515 | 2.511(13) | 2.506(24) |

**These target values were derived using free variables in DFIX instructions, taking advantage of the 4 and 8-fold redundancy and the high resolution (0.92Å) of the cytochrome $c_6$ structure.**

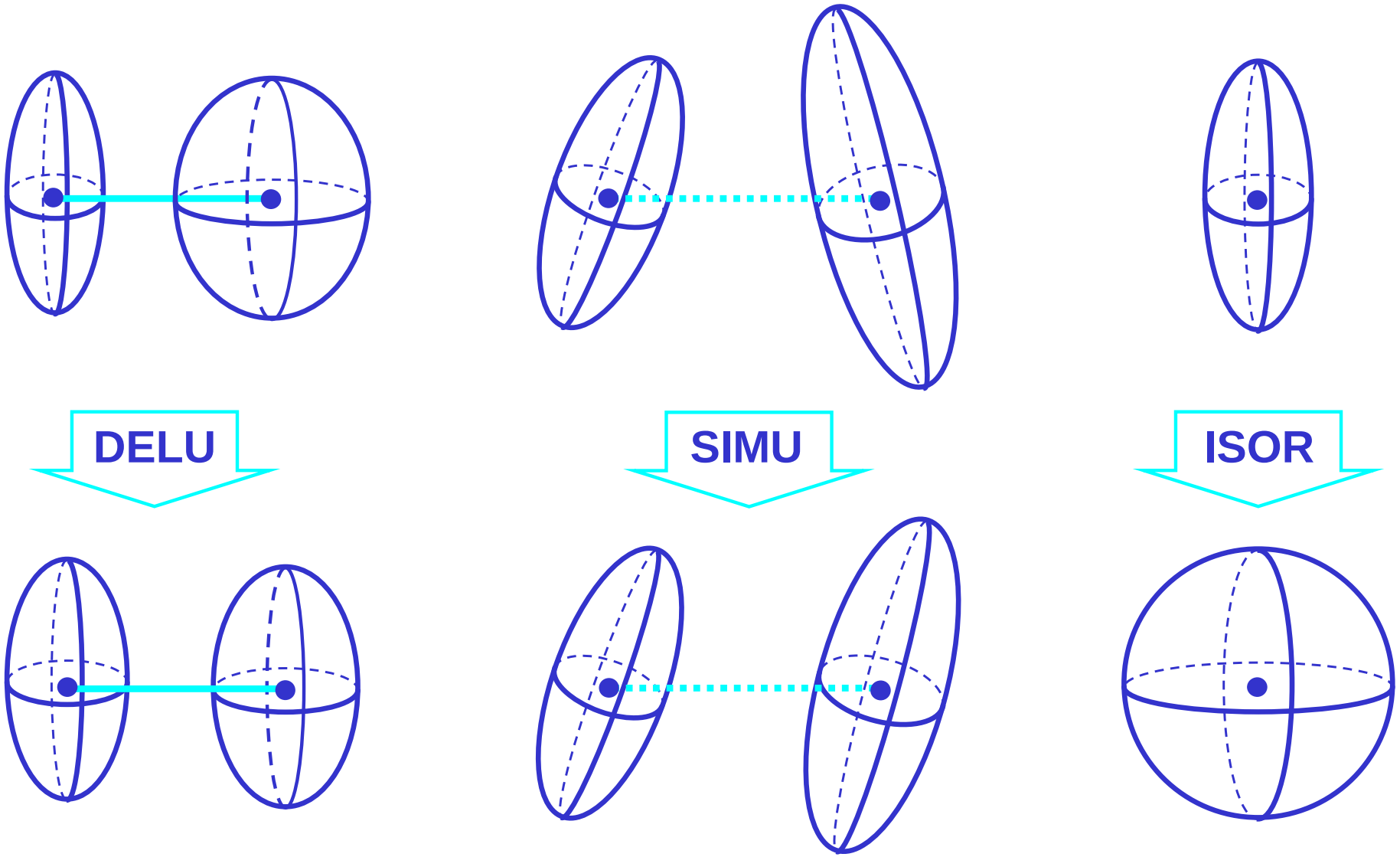
# Chiral volume and planarity restraints

**CHIV restrains the *chiral volume* of an atom that makes 3 bonds (ignoring H). The chiral volume is the volume of the 'unit-cell' (i.e. parallelopiped) with axes formed by the 3 bonds, and its sign is determined by the alphabetical order of the 3 atoms. This can be used to fix the chirality of $C_\alpha$ and (Ile and Thr) $C_\beta$. It can also be used to ensure that $C_{\gamma 1}$ and $C_{\gamma 2}$ of Val are labeled conventionally!**

```
CHIV_VAL  CA  2.516
CHIV_VAL  CB  –2.622
CHIV_VAL  C
```

**The last line restrains the chiral volume of the carbonyl C to zero (the default value) and is a convenient local planarity restraint. The program uses the connectivity array to find the three bonds. The FLAT planarity restraint is suitable for more general cases, e.g. the five atoms in a peptide plane or the 9 coplanar atoms of a Trp side-chain or a base in an oligonucleotide.**

# Restraints on ADP's




Thomas R. Schneider

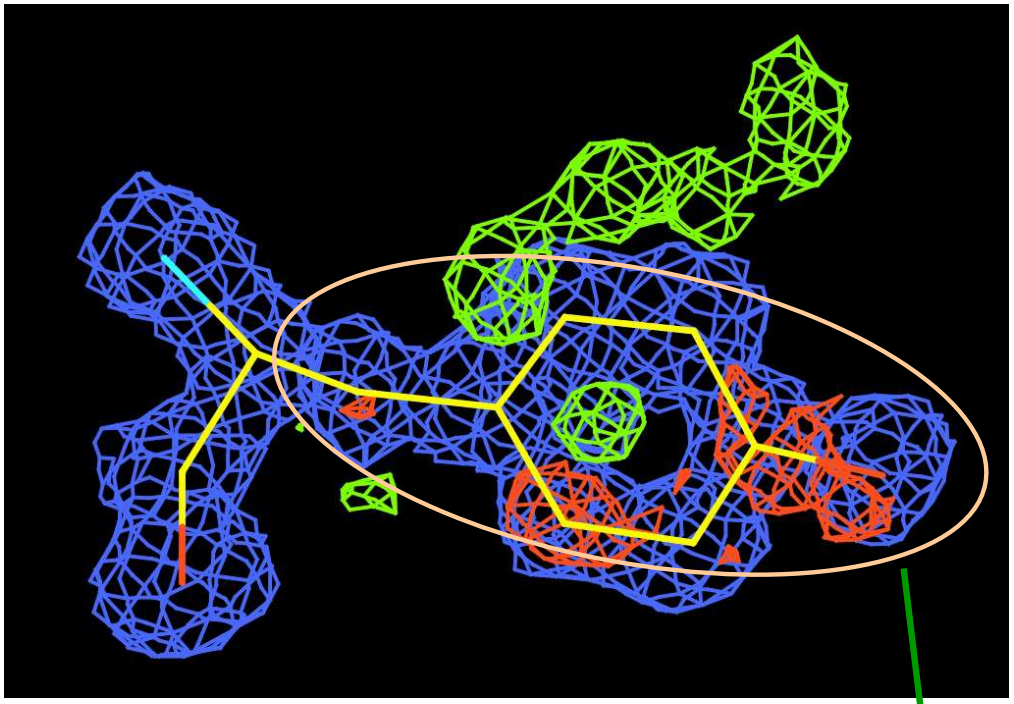
# NCS (non-crystallographic symmetry) restraints

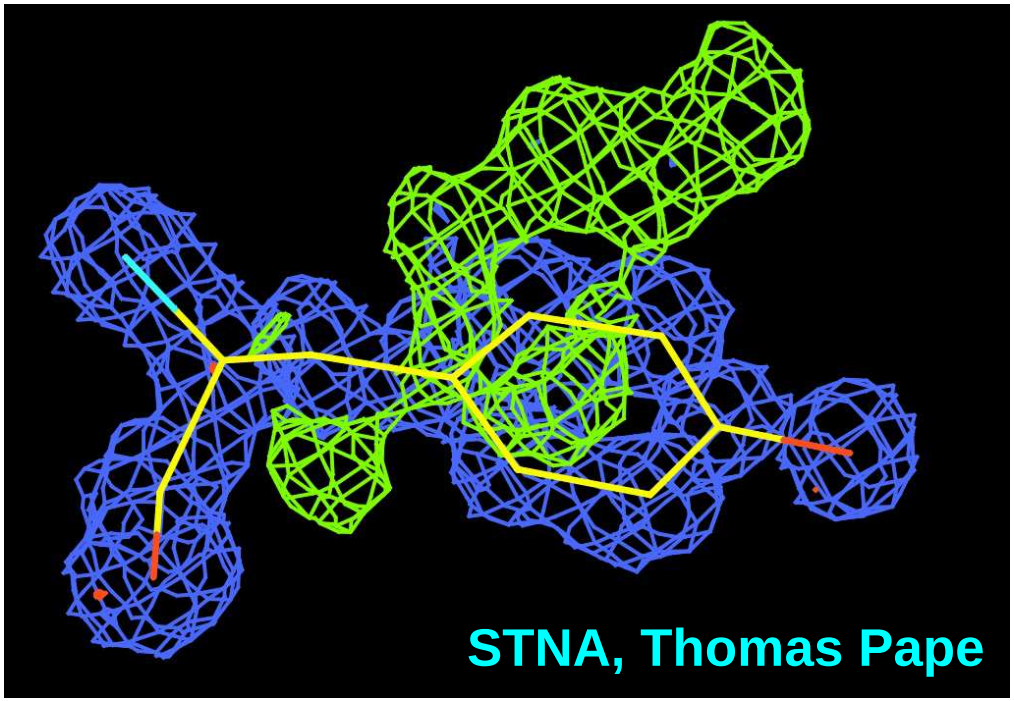
**SHELXL applies NCS as a *local restraint* rather than a *global constraint*. This is slower but more flexible and does not require a mask or transformation matrix.**

**The NCS-related 1,4-distances are restrained to be equal; this is similar to restraining torsion angles to be equal, but does not distinguish between ± gauche positions. In addition the isotropic *U*-values of NCS-related atoms may be restrained to be equal.**

```
NCSY 1000 N_1001 > OT2_1109
NCSY 2000 N_1001 > OT2_1109
```

**would specify threefold NCS, where the three chains are numbered 1001-1109, 2001-2109 and 3001-3109.**

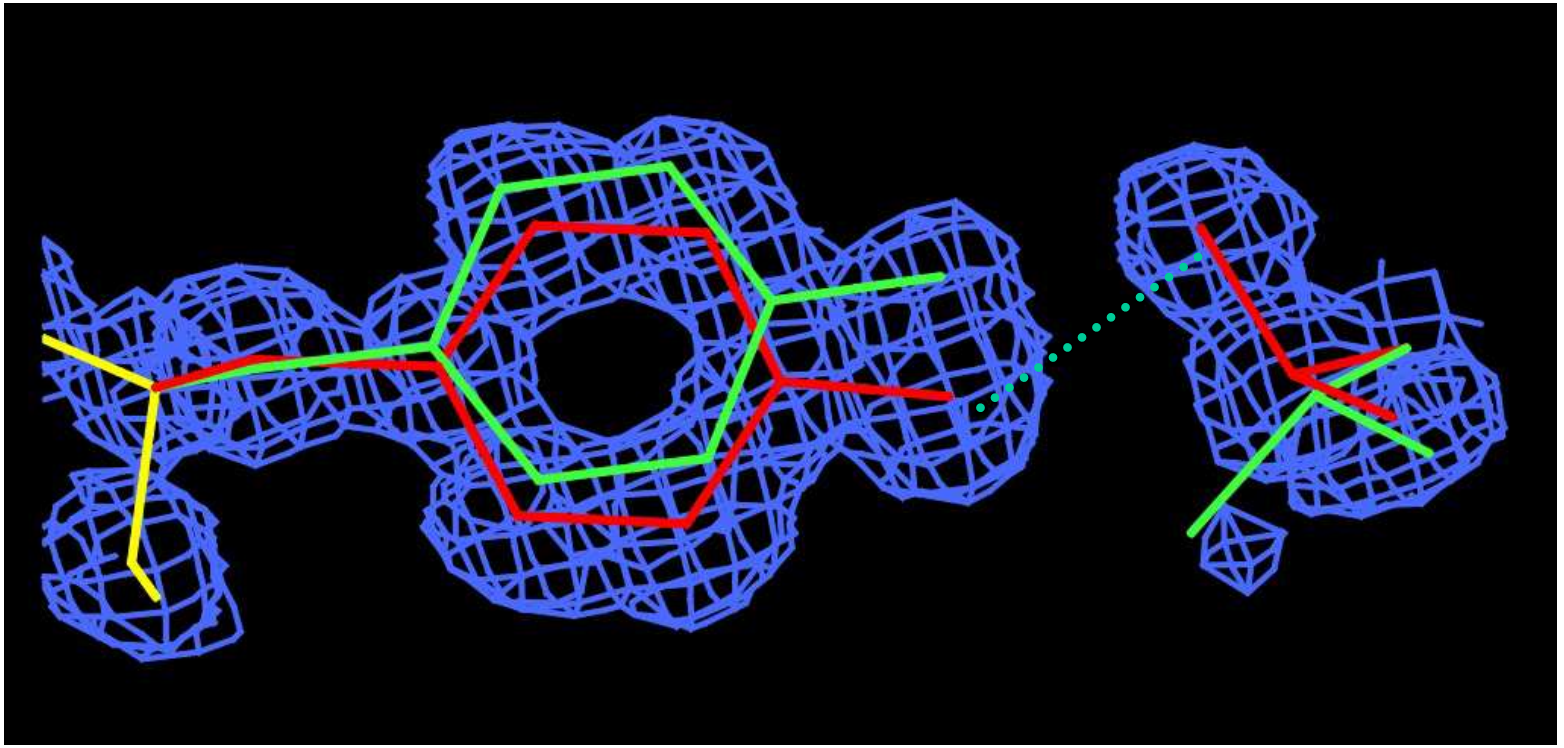





```
RESI  381   TYR
N     3  x y z 11.0 u11 u22 ..
CA    1  x y z 11.0 u11 u22 ..
CB    1  x y z 11.0 u11 u22 ..
CG    1  x y z 11.0 u11 u22 ..
..
OH    4  x y z 11.0 u11 u22 ..
C     1  x y z 11.0 u11 u22 ..
O     4  x y z 11.0 u11 u22 ..
```

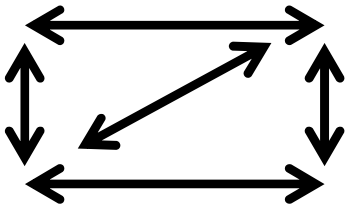
```
RESI  381   TYR
N     3  x y z 11.0 u11 u22 ..
CA    1  x y z 11.0 u11 u22 ..
PART 1 10.65
CB    1  x y z 11.0 u11 u22 ..
CG    1  x y z 11.0 u11 u22 ..
..
OH    4  x y z 11.0 u11 u22 ..
PART 0
C     1  x y z 11.0 u11 u22 ..
O     4  x y z 11.0 u11 u22 ..
```

Thomas R. Schneider



1 parameter describes the occupancies of 22 atoms !!

```
RESI 233 TYR
..
PART 1 31.0
CB   ..
PART 2 -31.0
CB   ..
PART 0
..
```



```
RESI 123 THR
..
PART 1 31.0
CB   ..
PART 2 -31.0
CB   ..
PART 0
..
```

Thomas R. Schneider

# Acknowledgements

**I am very grateful to Regine Herbst-Irmer, Thomas R. Schneider, Isabel Usón and Peter Müller for their many contributions to the development of refinement methods for high-resolution data, and to the rest of my group in Göttingen and all the SHELX users around the world for their comments and suggestions.**

**Standard SHELX refererence:**

**Sheldrick, G.M. (2008). *Acta Cryst*. A64, 112-122.**

**SHELXL book:**

**Müller, P., Herbst-Irmer, R., Spek, A., Schneider, T.R. & Sawaya, M.R. (2006). *Crystal Structure Refinement: A crystallographer's guide to SHELXL.* IUCr/Oxford University Press.**

**SHELXL for macromolecular refinement:**

**Schneider, T.R. & Sheldrick, G.M. (1997). *Methods Enzymol*. 277, 319-341.**