Rietveld Structure Refinement of Protein Powder Diffraction Data using GSAS

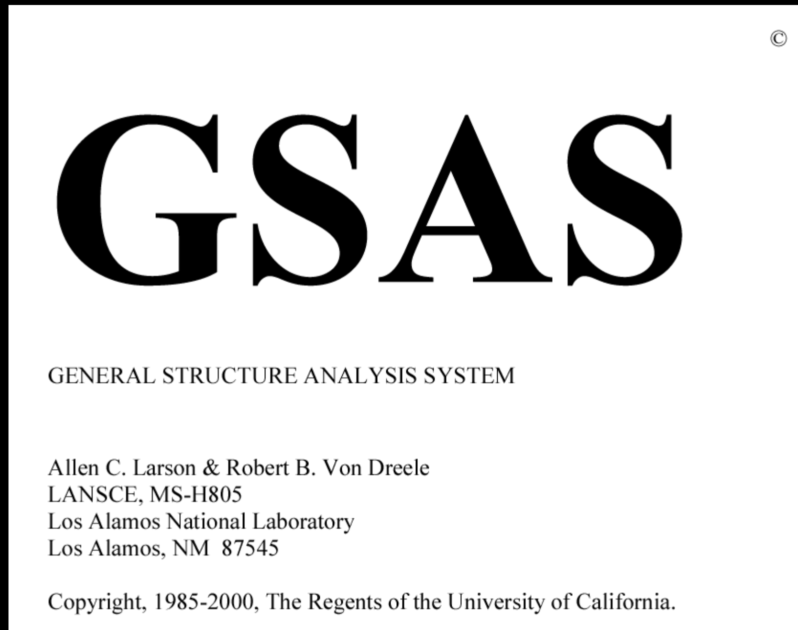
Jon Wright

ESRF, Grenoble, France

Plan…

- This is a users perspective
- Cover the protein specific aspects (assuming knowledge of Rietveld and some familiarity with GSAS)
- Reading in data/model and setting up a refinement
- Reading in restraints files, setting parameters for Marquardt damping & band matrix for least squares
- Use of spdbv for viewing structure, maps, modifying/fixing side chains, Ramachandran plots

A users perspective!

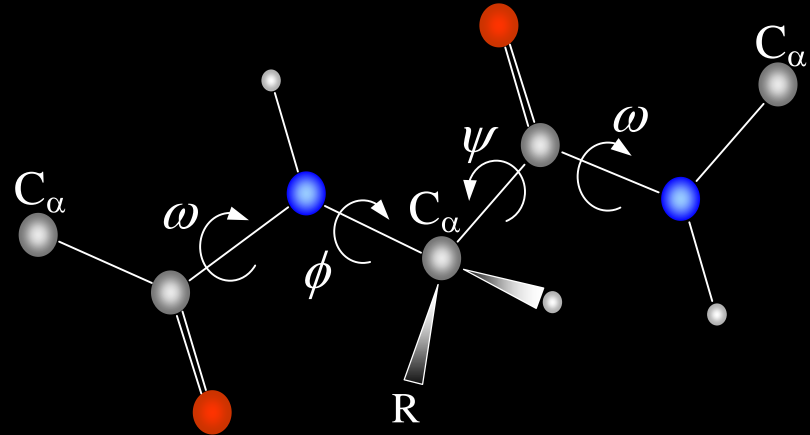


- GSAS – General Structure Analysis System, Larson and Von Dreele
- July 6,1999 – First "protein" version became available

Any errors, omissions and fundamental misunderstandings presented here should be attributed to me, and not GSAS or it's authors!

Protein structure from powder data??

- Fundamental differences to small molecule crystallography
  - Lots of geometric constraints
  - only ϕ, ψ and sidechain conformations which are unknown
- Low resolution crystallography



Restraints provide short range info, powder data for longer range

R. B. Von Dreele, *J. Appl. Cryst.* (1999). **32**, 1084-1089

"…refinement of the 1261-atom protein metmyoglobin was achieved by combining 5338 stereochemical restraints with a 4648-step (d_{min} = 3.3 Å) "

[3783 x,y,z co-ordinates from only 4648 datapoints!]

Setting up a refinement…

- Need a dataset and a PDB file containing a model
- Insert phase and flag as "macromolecule"
- Insert data as normal
- You might try a LeBail fit to get good peakshape and cell parameters

```
Enter phase edit command(<?>,$,D,E,F,M,I,L,R,S,X> >m 1
The phase is non-magnetic
Enter phase type (<?>,A,B,C,D,L,X> >
Selection of phase type:
 A  - Nuclear structure only
 B  - Nuclear and magnetic structure
 C  - Magnetic structure only
 D  - Macromolecular structure
Enter phase type (<?>,A,B,C,D,L,X> >d
The phase is for a macromolecule
```

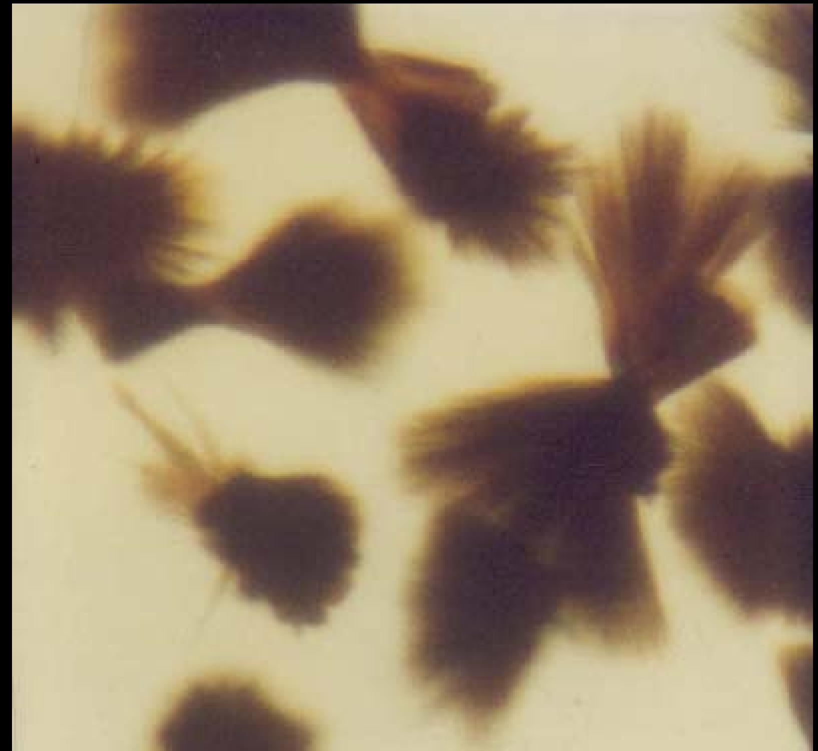
Tip!

For starting a LeBail fit only use the low angle portion of the pattern

(severe overlaps)

Example used here...

- Myoglobin (horse)
- Crystallised by salting out with ammonium sulphate
- Data collected at BM16, ESRF
- $\lambda = 0.4135Å$



See Jeremy Cockcroft (**FA 5 - MS 7**, Wednesday pm) for other examples at this conference

```
Phase No. 1; Phase has      0 atoms; Title: Myoglobin
Give atom editing command (<?>,$,I,S,X) >i
Command structure for inserting an atom
 I s - enter atom with sequence number "s"
 I N - enter atom with next sequence number
 I B - Read atoms from BNL PDB format
 I R - Read atoms from non-GSAS file or from another EXP file
Phase No. 1; Phase has      0 atoms; Title: Myoglobin
Give atom editing command (<?>,$,I,S,X) >i b
 Enter non-GSAS PDB input file name (<?>,$,QUIT)
  >pdb1ymb.ent
The first ten lines are:
HEADER    OXYGEN TRANSPORT                        27-SEP-93   1YMB      1YMB   2
COMPND    METMYOGLOBIN (HORSE HEART)                                    1YMB   3
SOURCE    HORSE (EQUUS CABALLUS)                                        1YMB   4
AUTHOR    S.V.EVANS,G.D.BRAYER                                          1YMB   5
REVDAT   1   31-JAN-94 1YMB    0                                        1YMB   6
JRNL        AUTH   S.V.EVANS,G.D.BRAYER                                 1YMB   7
JRNL        TITL   HIGH RESOLUTION STUDY OF THE THREE-DIMENSIONAL       1YMB   8
JRNL        TITL 2 STRUCTURE OF HORSE HEART METMYOGLOBIN                1YMB   9
JRNL        REF    J.MOL.BIOL.                    V. 213   885 1990     1YMB  10
JRNL        REFN   ASTM JMOBAK  UK ISSN 0022-2836               070     1YMB  11
Is this the correct file (<Y>/N/Q)? >y
Do you want to copy HOH molecules (Y/<N>)? >n
Do you want to copy disordered atoms (Y/<N>)? >n
Do you want to copy hydrogen atoms (Y/<N>)? >n
Select atom transform method (C=cell,<M>=matrix,N=none) >m
Phase No. 1; Phase has   1247 atoms; Title: Myoglobin
Give atom editing command
 (<?>,$,C,D,E,F,I,K,L,S,T,V,X,+,-,*,/) >
```

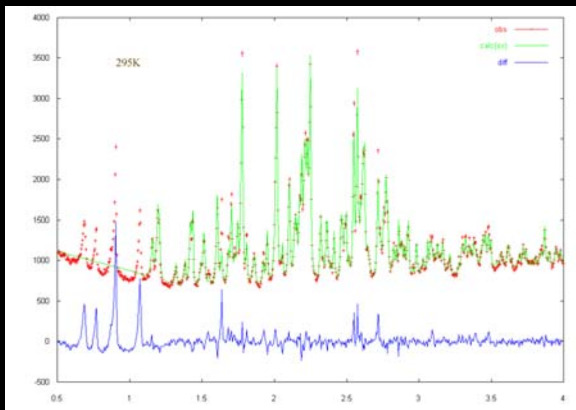
Inserting the atoms is trivial, when the structure is available in a PDB file !

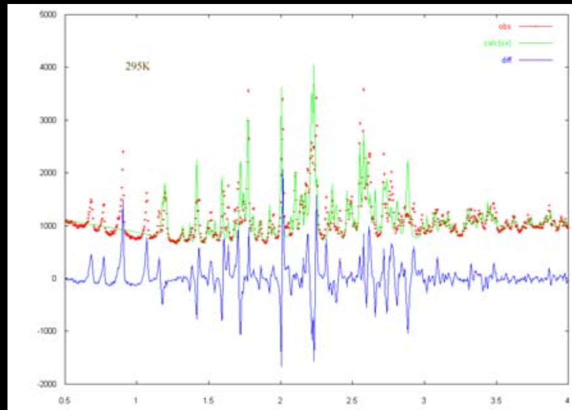
The PDB file, unit cell and dataset must “match”

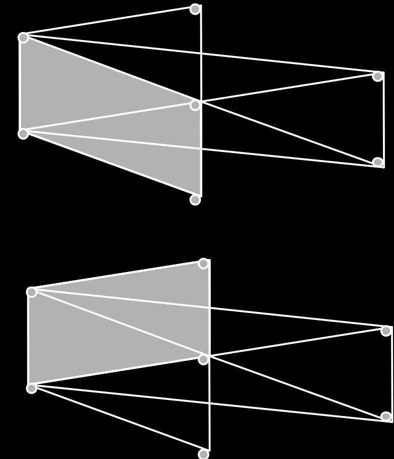
Choose the right cell

- Indexing produced two (equivalent) unit cells, which correspond to two choices for a monoclinic set of axes, both close to an orthorhombic cell.

	Cell 1	Cell 2
a =	64.20502	63.5722
b =	28.92236	28.9320
c =	35.85339	35.8645
beta =	107.1616	105.386
Chi^2 =	21.28	122.92



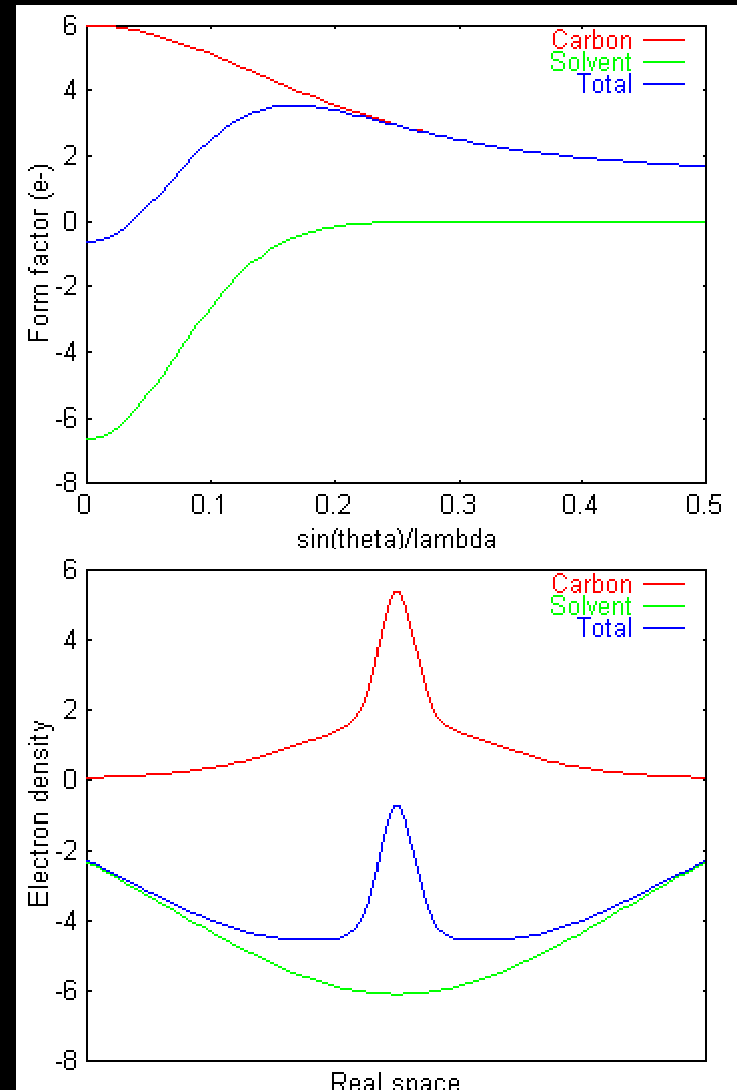




Fits using deposited single crystal structure factors (works better than comparing to the model !!)

Solvent scattering…

- Tends to build up electron density in regions which are far from any atoms
- Fills in the void space in the structure with smooth density
- Works "surprisingly well"
- Parameters have some physical meaning
- Expedt: L F S

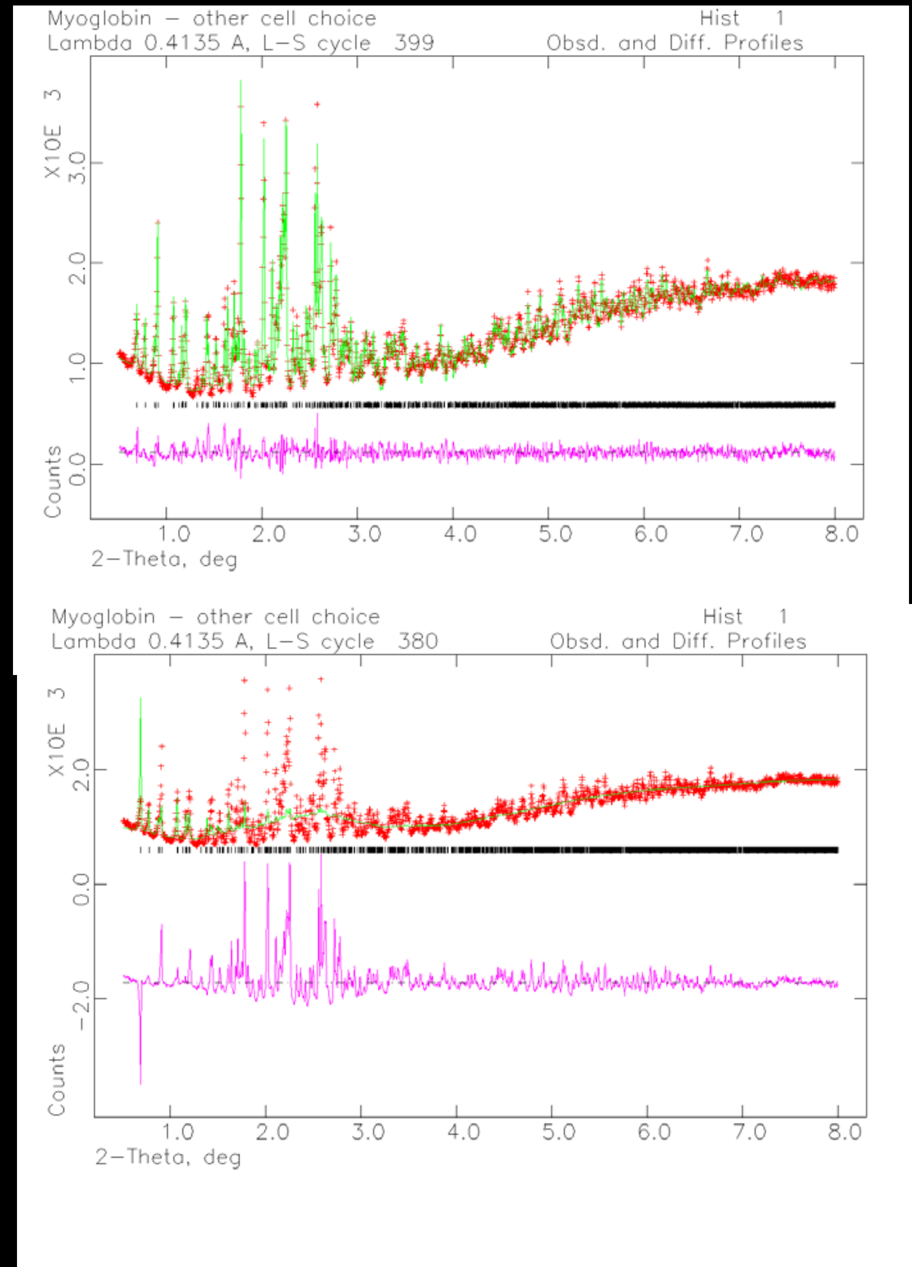


Solvent scattering

- Fits with and without solvent contribution
- "Babinet's principle" gives modified atomic scattering factors:

$$f = f_0 - A\exp\left(\frac{-8\pi^2 U \sin^2\theta}{\lambda^2}\right)$$

- Here $A=6.63e^-$ and $U=1.18Å^2$ (so carbon effectively has no electrons at low angles!)



Setting up restraints:

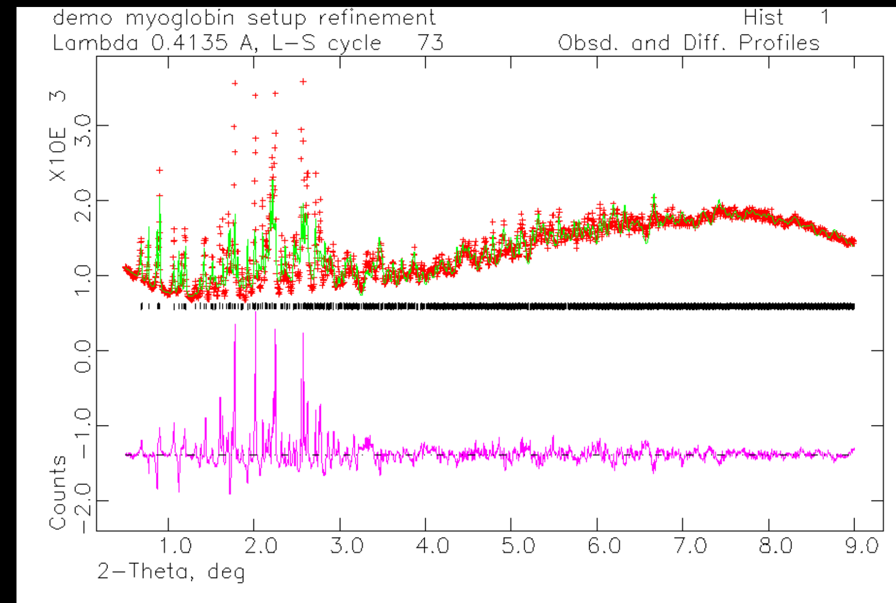
- GSAS provides macros containing standard amino acid geometries
- Macromolecule phases identify amino acids for working out restraints
- Treat non-amino acid groups manually (eg heme, ligands etc)

Use the macro files in $GSAS/macros/, then
"expedt, L S" for soft restraints
a @r angles.mac
x d @r bonds.mac
x k @r chiral.mac
x r @r rama.mac
x t @r torsion.mac
x p @r planes.mac
…takes a few minutes, depending on the speed of the computer (and typist)

```
Phase No. 1; Phase has   1247 atoms; Title: Myoglobin 64Angstrom cell
  SER TYPE      X       Y       Z      FRAC      UISO NAME RES   NO GRP  CODE FXU
    1 N     0.02858 0.57271 0.42585 1.00000 0.30000 N    GLY    1              000
    2 C     0.02064 0.53126 0.44021 1.00000 0.30000 CA   GLY    1              000
    3 C     0.03894 0.50304 0.46720 1.00000 0.30000 C    GLY    1              000
    4 O     0.03648 0.46117 0.47118 1.00000 0.30000 O    GLY    1              000
    5 N     0.05741 0.52634 0.48595 1.00000 0.30000 N    LEU    2              000
    6 C     0.07666 0.50396 0.51271 1.00000 0.30000 CA   LEU    2              000
    7 C     0.07751 0.50598 0.55618 1.00000 0.30000 C    LEU    2              000
    8 O     0.07490 0.54254 0.57189 1.00000 0.30000 O    LEU    2              000
```

Get a good fit BEFORE refining atom positions!

- Time spent ensuring the peak shape, unit cell and background are good will pay off later
- Try at least one cycle with restraints present but positions fixed to see how well the geometry matches



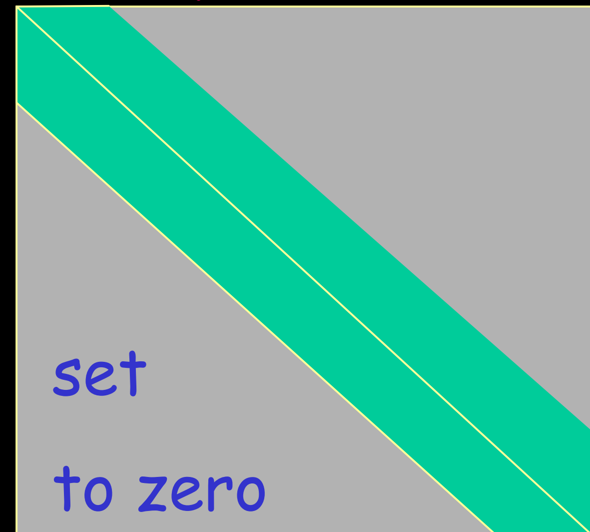
The restraints should be well fitted at the start if you have a sensible model

Cross your fingers....

- And let all the co-ordinates go!
- Don't refine UISO, just set it to 0.3.
- Apply some Marquardt damping (stability)
- Set max atom shift to 0.25-0.5Å
- Set the matrix bandwidth (speed)
  - try to be wide enough to preserve restraints

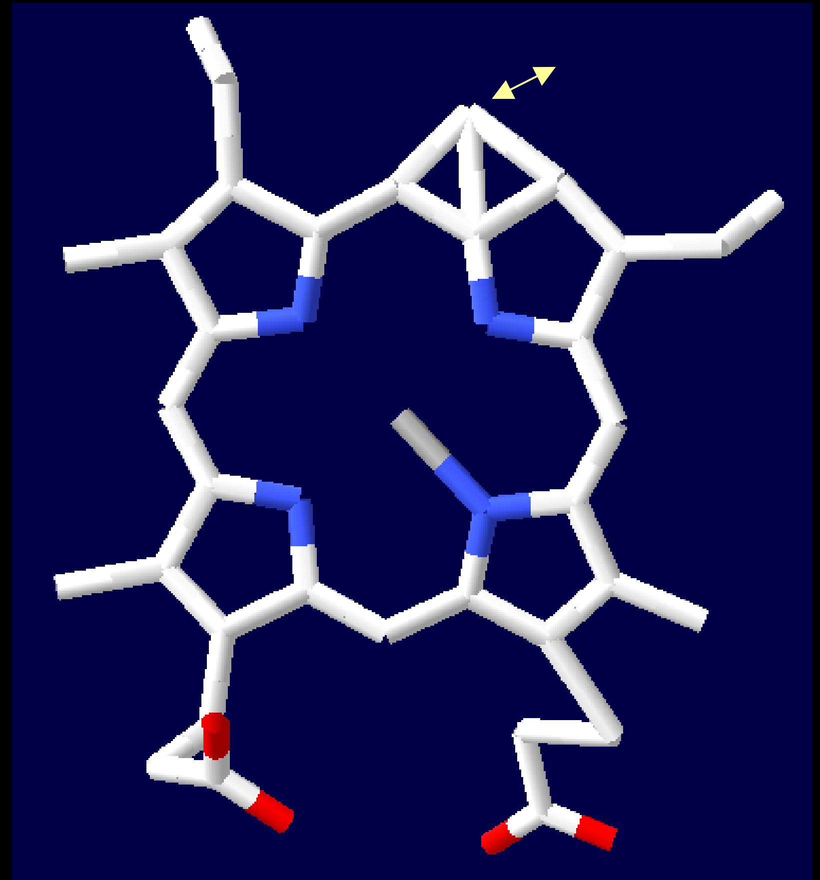
Damping is the restraint/observation that the current model is correct

Multiplies diagonal of LSQ matrix by Marquardt factor

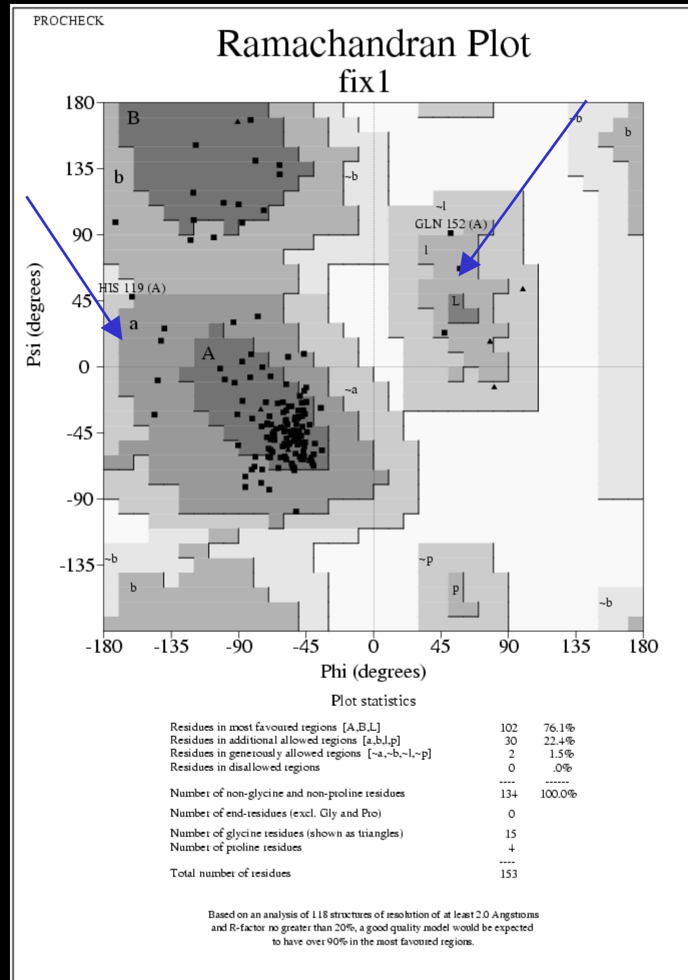


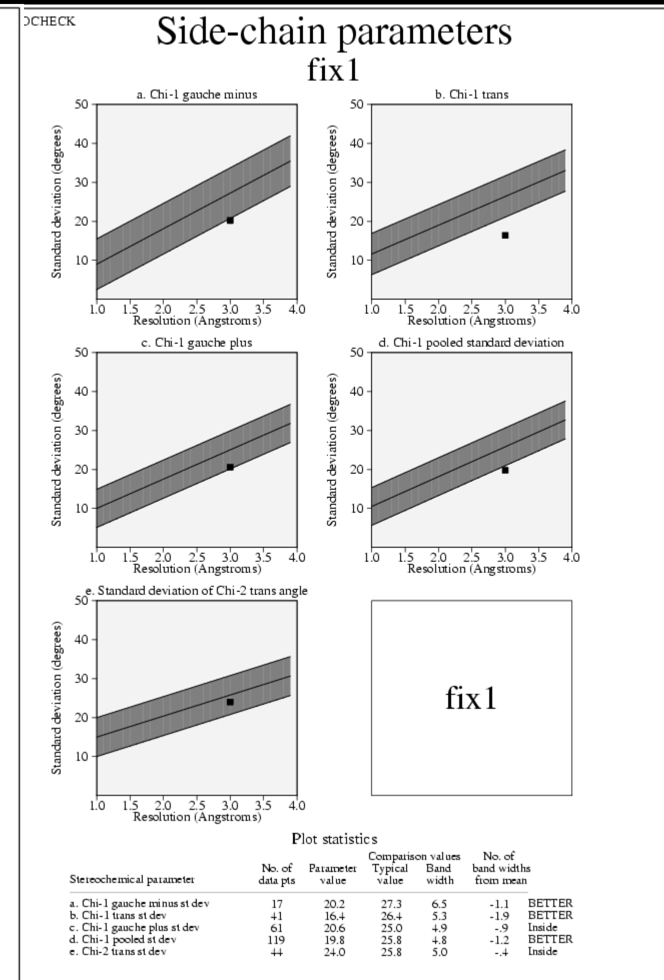
Did it fly to pieces?

- Use the "GSAS2PDB" program to output your structure
- Use some more specialised protein software to inspect your refined structure



- Detailed geometry checks for ensuring a sensible structure
- Highlights areas needing attention





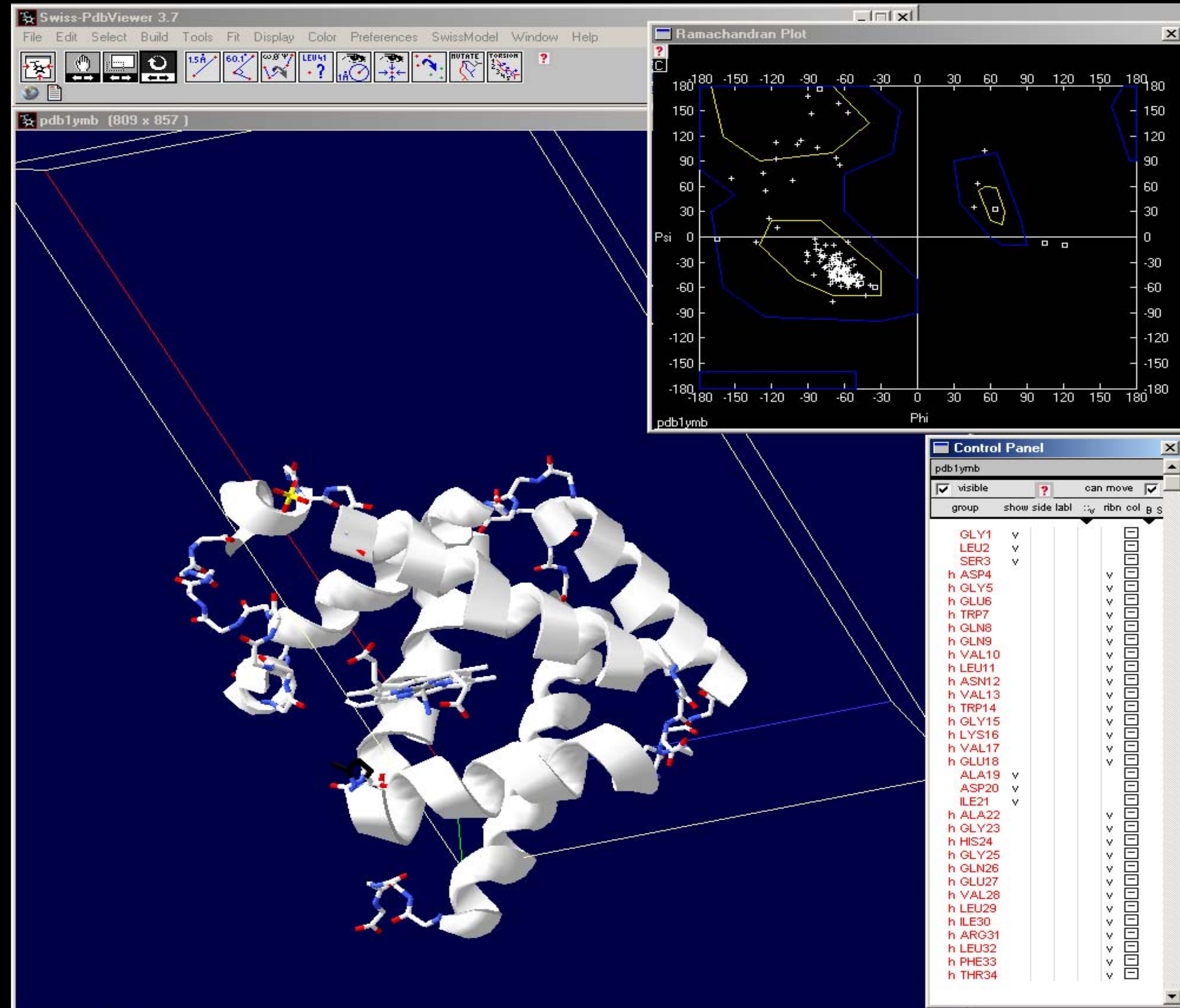
PROCHECK

Laskowski R A, MacArthur M W, Moss D S & Thornton J M (1993). “PROCHECK: a program to check the stereochemical quality of protein structures.” *J. Appl. Cryst.*, **26**, 283-291.

http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html

spdbv : an ideal partner to GSAS

- FORPLOT writes fourier maps which can be read here
- Reads and writes pdb files
- Allows side chains to be "repaired"



freeware from http://www.expasy.org/spdbv

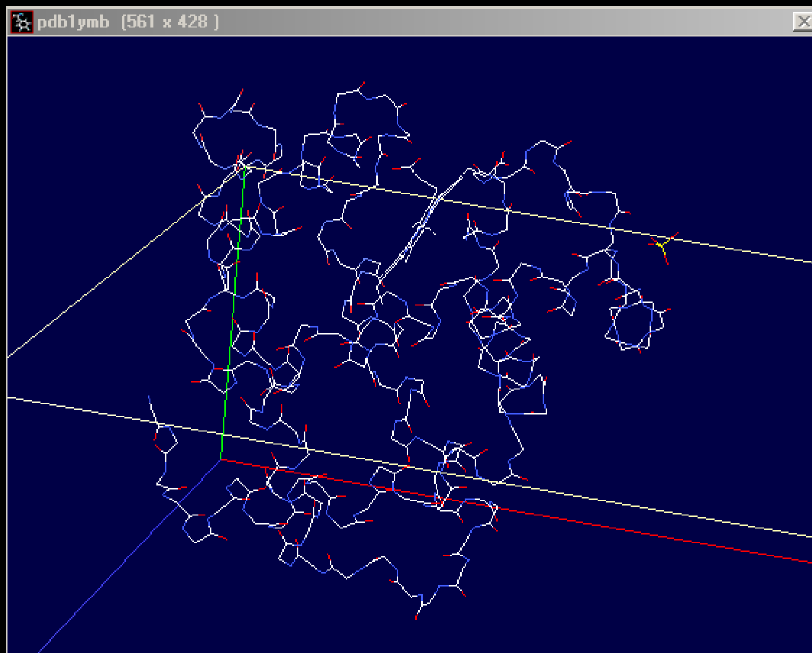
C:\spdbv\temp\pdb1ymb.E1

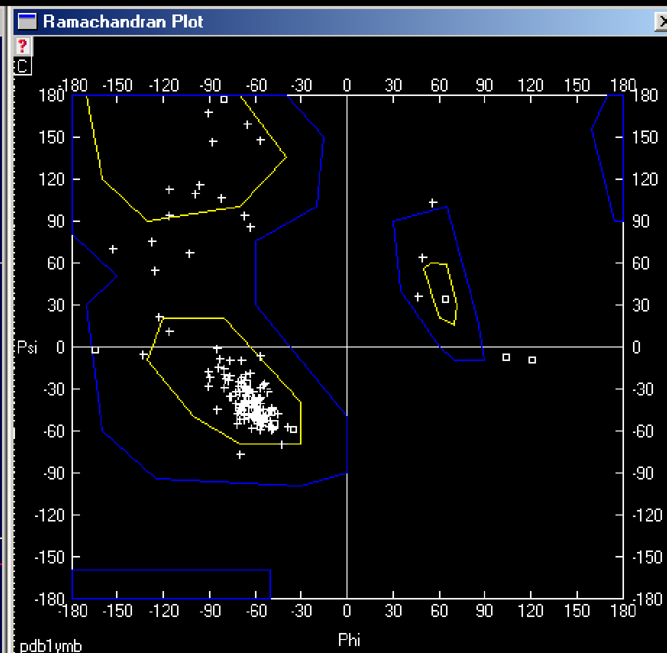
/ Computations were done in vacuo with the GROMOS96 43B1 parameters set, without reaction field.
/ For more information about GROMOS96, refer to: W.F. van Gunsteren et al. (1996) in Biomolecular
/ simulation: the GROMOS96 manual and user guide. Vdf Hochschulverlag ETHZ (http://igc.ethz.ch/gromos).
/ When using those results, please mention that energy computations were done with the GROMOS96
/ implementation of Swiss-PdbViewer.

/ residue		bonds	angles	torsion	improper	nonBonded	electrostatic	constraint //		TOTAL
HHT	1	0.000	6.183	7.540	0.000	0.00	22.26	0.0000 //	E=	35.980
GLY	1	6.421	4.222	0.007	0.109	-8.65	115.82	0.0000 //	E=	117.923
LEU	2	6.065	10.230	6.001	1.243	-26.66	11.71	0.0000 //	E=	8.588
SER	3	2.188	4.948	7.436	0.363	-6.30	-23.47	0.0000 //	E=	-14.843
ASP	4	9.217	17.925	4.083	4.254	9.98	50.42	0.0000 //	E=	95.873
GLY	5	9.265	1.518	3.873	2.971	-4.78	38.80	0.0000 //	E=	51.644
GLU	6	13.956	14.105	4.403	0.197	-27.36	-2.70	0.0000 //	E=	2.600
TRP	7	37.993	19.004	5.438	5.254	-65.50	-26.27	0.0000 //	E=	-24.082
GLN	8	10.526	12.064	5.452	3.059	-11.42	-179.31	0.0000 //	E=	-159.630
GLN	9	10.937	15.708	21.399	0.779	-34.66	-178.93	0.0000 //	E=	-164.761
VAL	10	11.314	16.429	1.336	5.523	-37.90	-6.86	0.0000 //	E=	-10.163
LEU	11	17.793	15.757	6.296	6.248	-33.19	-3.11	0.0000 //	E=	9.799
ASN	12	8.735	13.702	4.279	4.505	-13.55	-161.76	0.0000 //	E=	-144.087
VAL	13	4.801	4.665	0.618	3.691	-8.40	-9.84	0.0000 //	E=	-4.470
TRP	14	15.196	36.765	3.005	15.532	-58.04	9.12	0.0000 //	E=	21.574
GLY	15	7.917	6.685	0.212	0.449	-5.32	27.15	0.0000 //	E=	37.096
LYSH	16	11.320	7.049	23.778	0.112	-37.61	0.06	0.0000 //	E=	4.714
VAL	17	15.933	14.665	2.338	1.384	-33.52	-9.97	0.0000 //	E=	-9.167
GLU	18	15.664	19.363	8.251	1.098	-24.30	-3.05	0.0000 //	E=	17.032
ALA	19	15.024	4.169	0.383	3.589	-10.73	-1.54	0.0000 //	E=	10.897
ASP	20	7.953	7.522	4.062	0.454	-21.29	-19.69	0.0000 //	E=	-20.996
ILE	21	20.314	10.581	3.630	3.783	-9.34	-8.14	0.0000 //	E=	20.822
ALA	22	10.590	19.177	0.694	8.483	-12.69	36.98	0.0000 //	E=	63.230
GLY	23	1.792	6.117	1.570	0.064	-11.06	42.73	0.0000 //	E=	41.206

Structure energy calculations in spdbv complement the restraints used in GSAS to highlight any problematic areas

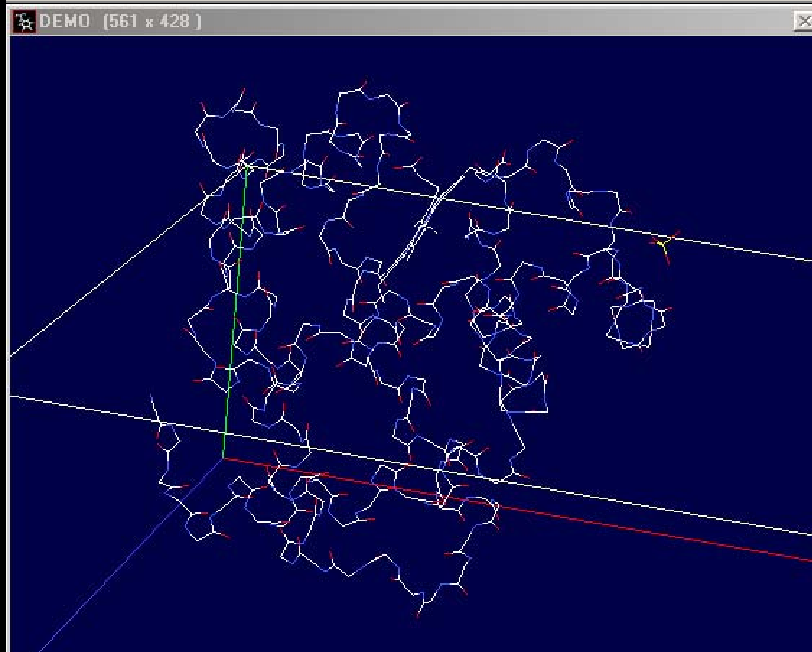
(GROMOS, inside spdbv ignores some intermolecular non-bonded contacts which GSAS includes, so they don't exactly agree)

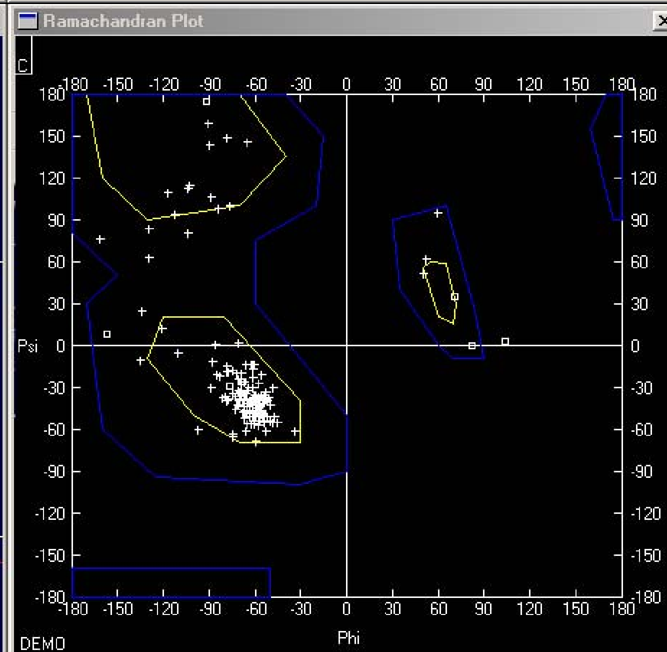




Rwp

10.18%





Rwp

7.65%

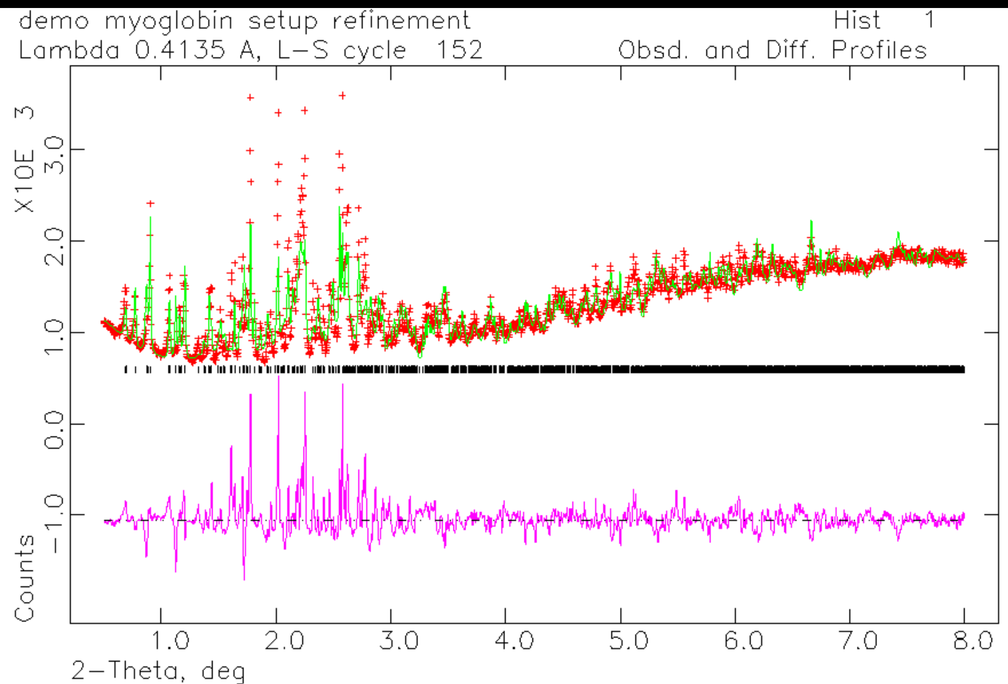
Rwp ~ 10.18 %

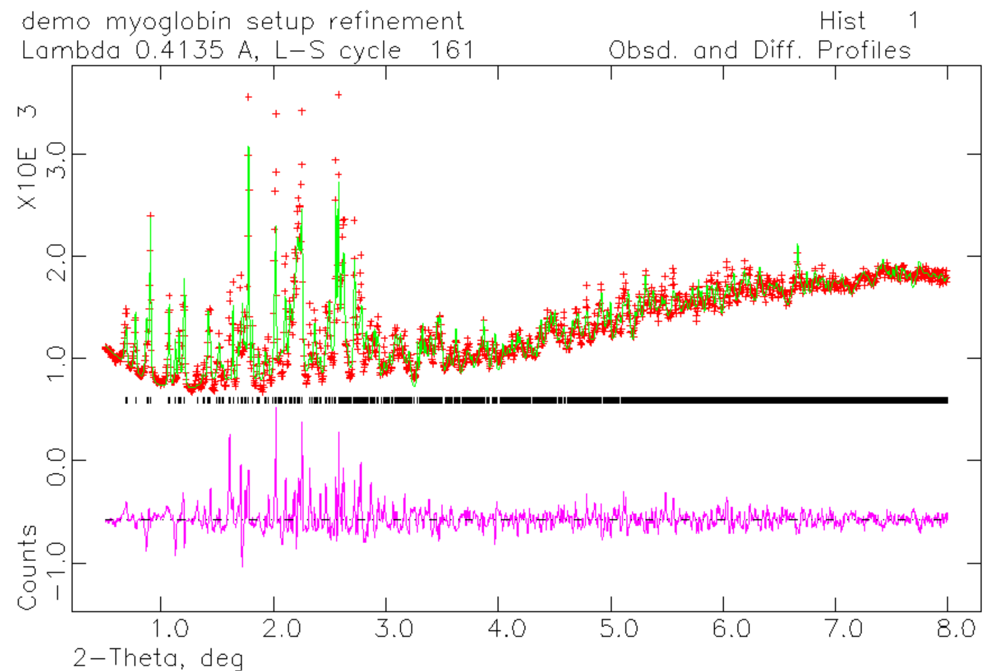
~10 cycles

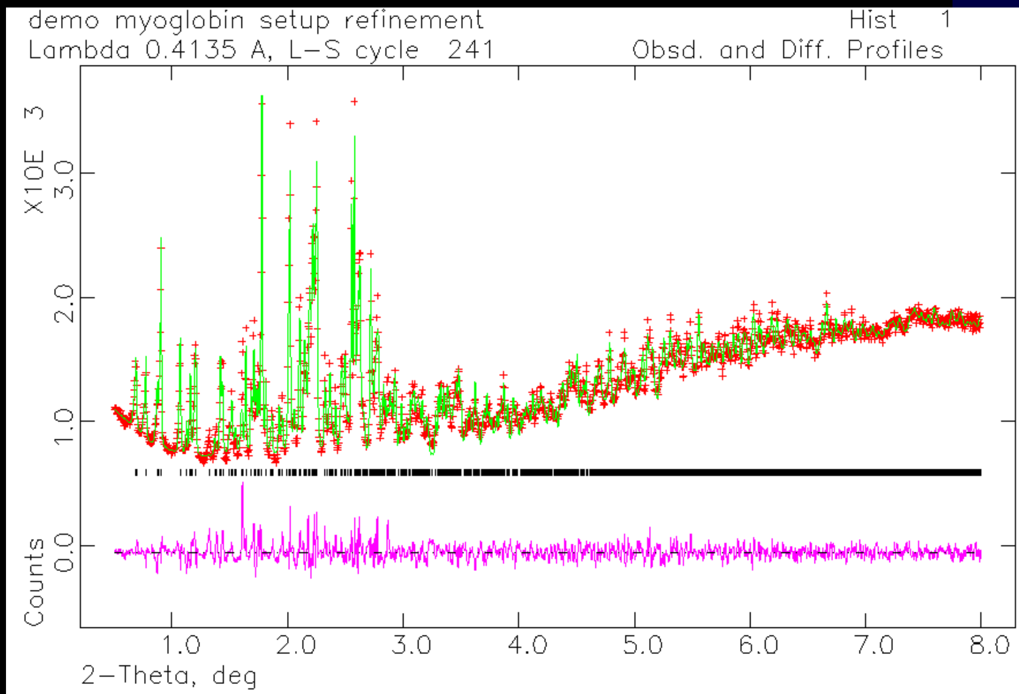
takes around 2 minutes per cycle (1.6 GHz P4)

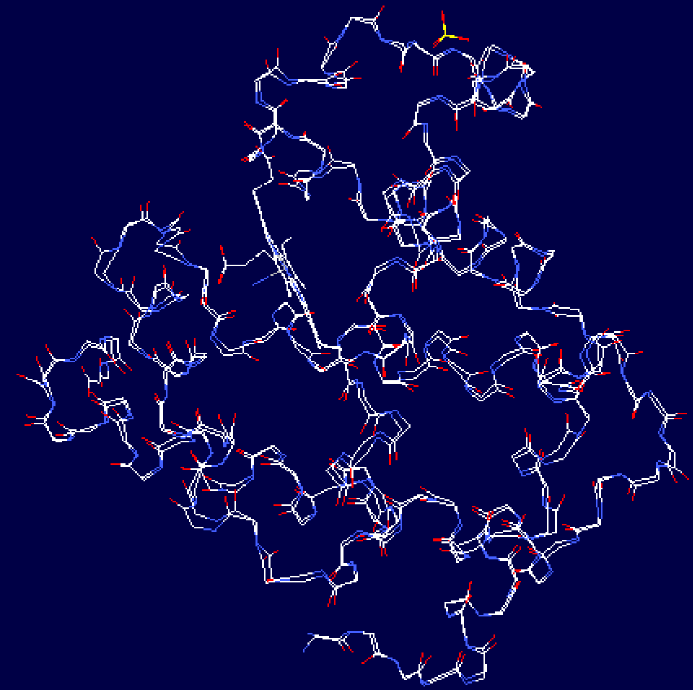
Not even time for lunch!

Rwp ~ 7.65 %







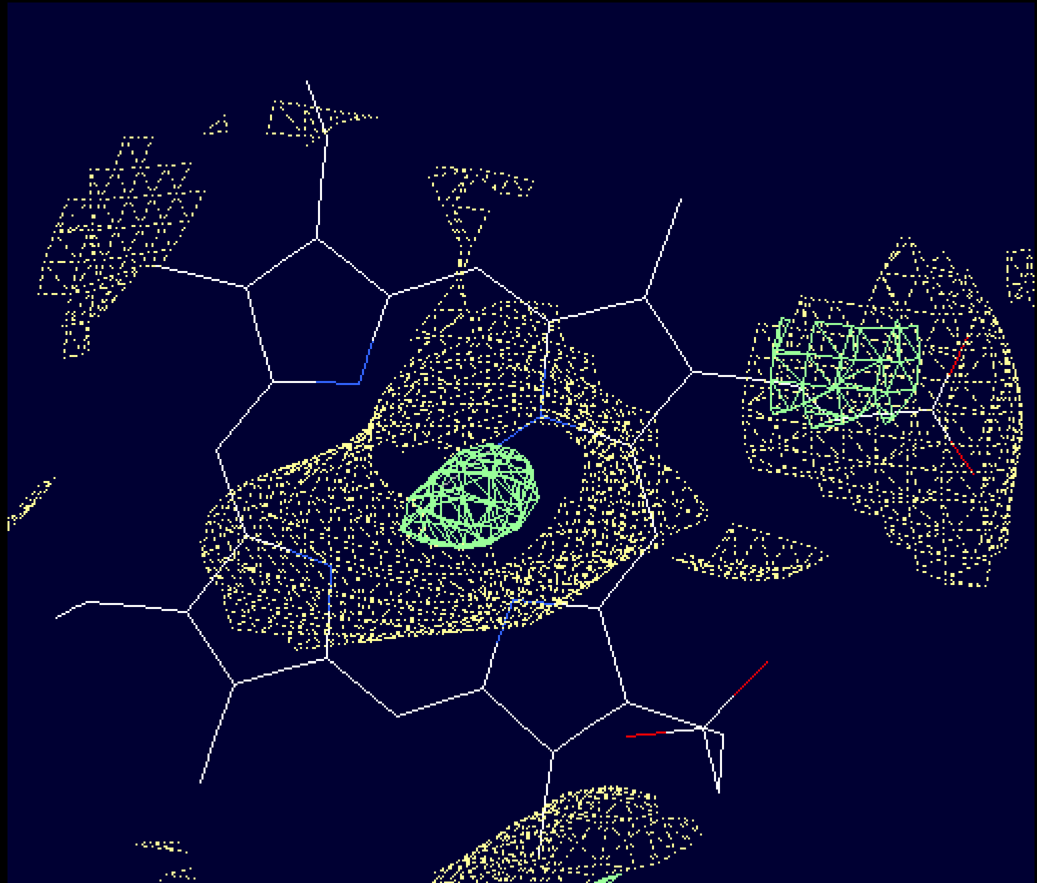


(another 60 cycles later....) Rwp 4.53%

Further improvement is doubtless possible, but the structure essentially reproduces the data and still gives about the same folds as the original. Geometry is still (just) acceptable for the checks in spdbv.

A (biased) omit map

- Fix the x,y,z of everything
- Set the occupancy of the interesting region to zero
- Run a (zero cycle) genles + fourier
- FORPLOT for difference map
  - o option to make a DN6
- Read into spdbv along with pdb file to visualise



Summary

- Setting up restraints for a "standard" amino acid chain is automated
- Interaction with PX software is facilitated by PDB and DN6 map file formats
- New features can be applied elsewhere
  - Solvent scattering – zeolite channels?
  - Marquardt damping – difficult structures
  - DN6 maps for any fouriers
  - PDB files for easy H-bonding checks in mercury (visualisation software from CCDC)
  - Macro files for any restrained refinements

Don't forget...

- Flag the phase as a macromolecule
  - p p m 1 d
- Put some solvent scattering in
  - l f s
- Add and check the restraints
- Use a band matrix and some damping
  - l l b 50 (or more)
  - l l d 1.5
- Bob Von Dreele has deposited data and structures for refinements in IUCr journals – you can practise on these
- ccp14 tutorial at: http://www.ccp14.ac.uk/solution/gsas/peak_and_proteins.html

Acknowledgement

- Bob Von Dreele
  - Giving away his program
  - Lots of useful advice and encouragement

- ESRF
  - my generous employer!