Chemical Crystallography and Structural Chemistry

(VO 270287)

Lecture 8

28th May 2020

Dr. Tim Grüne
Centre for X-ray Structure Analysis
Faculty of Chemistry
University of Vienna

tim.gruene@univie.ac.at

Previous Lecture

1. Model building
2. Refinement
3. Constraints & Restraints

Today's Lecture

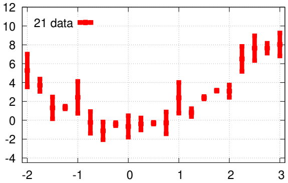
1. Example of constraints and data:parameter ratio
2. Validation

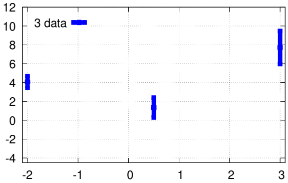
Example: Stabilisation through restraints

Two hypothetic measurements:

Experiment 1: high resolution, 21 pairs of measurements $(x_1, y_1), \dots, (x_{21}, y_{21})$ and errors $\sigma_1, \dots, \sigma_{21}$

Experiment 2: low resolution, 3 pairs of measurements $(x_1, y_1), \dots, (x_3, y_3)$ and errors $\sigma_1, \dots, \sigma_3$





Example: Stabilisation through restraints

Testing two models:

Model 1: $g(x) = g_2x^2 + g_1x + g_0$

Model 2: $h(x) = h_3x^3 + h_1x + h_0$

Either model has three parameters, g_0, g_1, g_2 and h_0, h_1, h_3 respectively. These parameters correspond to *e.g.* the model coordinates (x_i, y_i, z_i), or the ADPs U_i.

We will fit both models to the data to find out which model better represents the data.

Example: Stabilisation through restraints

Least-squares-minimisation:

$$\text{minimise: } \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(y_i - g(x_i))^2 \qquad \text{model 1}$$

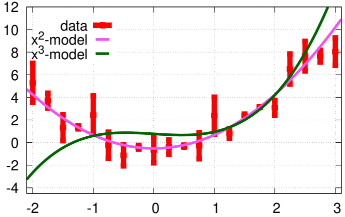
$$\text{minimise: } \sum_{i=1}^{N} \frac{1}{\sigma_i^2}(y_i - h(x_i))^2 \qquad \text{model 2}$$

- Experiment 1: $N = 21$ data points
- Experiment 2: $N = 3$ data points

We will start with the high resolution experiment 1

Example: Stabilisation through restraints

experiment 1: high resolution; high data to parameter ratio = 21:3=7



Model 1: $1.2x^2 + 0.0x - 0.5$

rmsd: 1.07

Model 2: $0.5x^3 - 0.3x - 0.8$

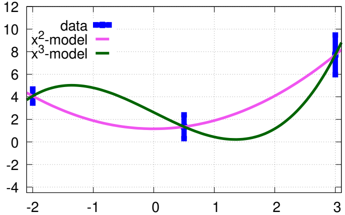
rmsd: 4.74

The *root mean square deviation* rmsd between model and data corresponds to the crystallographic $R1$ value.

The lower rmsd 1.07 clearly favours model 1. The pink curve also visually fits the data better than the green curve.

Example: Stabilisation through restraints

experiment 2: low resolution, low data to parameter ration = 3:3 = 1



model 1: $0.7x^2 + 0.0x + 1.2$

rmsd: 0

model 2: $0.5x^3 - 2.7x - 2.6$

rmsd: 0

When there are as many parameters as data points, any model can be fitted perfectly to the data. We cannot distinguish between the two models

Example: Stabilisation through restraints

experiment 2: low resolution with constraint

For some reason we know that the data must pass through the point $(0, 0)$. For the two models this means

$$\begin{aligned} 0 &= g(0) \\ &= g_2 * 0^2 + g_1 * 0 + g_0 \\ \Rightarrow g_0 &= 0 \end{aligned}$$

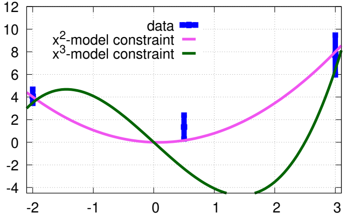
and analogously

$$h_0 = 0$$

The constraint reduced the number of parameters, only two parameters per model

Example: Stabilisation through restraints

experiment 2: low resolution with constraint



model 1: $0.9x^2 - 0.1x$

rmsd: 1.13

model 2: $0.8x^3 - 4.9x$

rmsd: 3.7

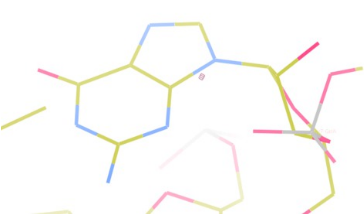
Due to the constraint, data to parameter ratio $=$ 3:2 $=$ 1.5. Now there is an *rmsd*, and it favours (again) the first model.

Summary & model building

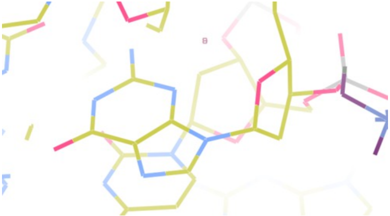
- (*cf.* phase problem)
- phases are calculate from the model
- model phases and observed data yield the electron density map, and electron difference map
- model building improves the model in large steps
- refinement optimises the model against the data
- medium resolution data or poor quality data require restraints and constraints in order to create a chemically sensible model

Model quality and data quality: structure validation

Atom coordinates $\neq$ model accuracy



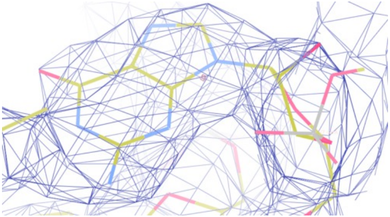
Guanine model in ribosome, data resolution 3.1Å



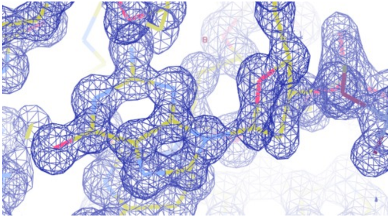
Guanine model in Z-DNA, at resolution 1.0 Å

The coordinates of the model do no reveal the data quality, nor the model quality.

model coordinates = interpretation of data



Guanine model **with map** in ribosome, data resolution 3.1Å



Guanine model **with map** in Z-DNA, at resolution 1.0 Å

Only in combination with the data can we judge the model quality

Once more: data to parameter ratio

Example Ciprofloxacin ($a = 9.5$Å, $b = 9.9$Å, $c = 11.0$Å, $\alpha = 94.2°$, $\beta = 100.2°$, $\gamma = 91.3°$)

- $FC_{17}N_3O_9H_{30}$: $60 \times 9 = 540$ Parameter

data resolution 0.43 Å: 26'308 reflections $\hat{=}$ 48.7 data points per parameter: very high, reliable refinement

data resolution 0.8 Å: 2'926 reflections $\hat{=}$ 5.4 data points per parameter: medium, refinement needs checking

Once more: data to parameter ratio

Example Ribosome ($a = 401.4$Å, $b = 401.4$Å, $c = 175.9$Å, $\alpha = \beta = \gamma = 90°$, $P4_12_12$)

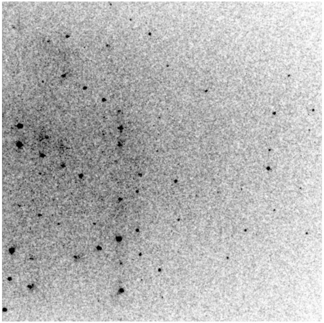
- PDB ID 1J5E: 51'atoms atoms = 207'768 parameters
- data resolution 3.05 Å 238'205 reflections

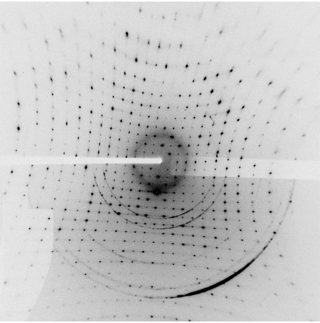
$$\frac{238'205}{207'768} = 1.15$$

Even at such low data to parameter ratio can a reasonable model be built and refined. It is important to be aware of differences in the interpretation of the data

Quality indicators

Example data quality





Important quality indicators

R$_{\text{meas}}$ relative difference between:

1. measured data
2. calculated data

data completeness : fraction of measured data w.r.t. theoretically possible data

multiplicity (*alias*: *redundancy*): how often every unique reflection was measured (on average)

signal strength $I(hkl)/\sigma_{I(hkl)} < 1$: noise

CC$_{1/2}$

1. split data set into two random halves
2. calculated correlation coefficient between symmetry equivalent reflections

R-values for data

The classic data quality indicator is $\mathrm{R_{int}}$, alias $\mathrm{R_{merge}}$ or $\mathrm{R_{sym}}$:

$$R_{\text{int}} = \sum_h \sum_j \frac{|I_{hj} - \langle I_h \rangle|}{\langle I_h \rangle}$$

$\mathrm{R_{int}}$ mathematically increases with multiplicity, although data quality improves with multiplicity

$\mathrm{R_{int}}$ is typically shown in publications. It is, however, obsolete and should not be published. $\mathrm{R_{meas}}$ *alias* $\mathrm{R_{r.i.m.}}$ should be published instead:

$$R_{\text{meas}} = \sum_h \frac{n_h}{n_h - 1} \sum_j \frac{|I_{hj} - \langle I_h \rangle|}{\langle I_h \rangle}$$

Example data statistics (XPREP)

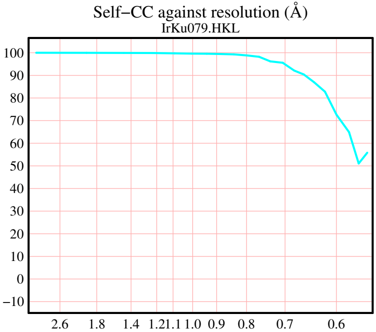
```
  Resolution  #Data #Theory %Complete Redundancy Mean I Mean I/s Rmerge
 Inf - 2.46     196    197     99.5    39.27    215.01  110.27  0.0300
2.46 - 1.13    1762   1825     96.5    14.86     75.32   42.01  0.0453
1.13 - 0.89    1972   2123     92.9     8.71     25.52   19.00  0.0895
0.89 - 0.77    2007   2258     88.9     6.81     10.84   10.39  0.1425
0.77 - 0.69    1864   2499     74.6     3.37      5.66    5.76  0.1885
0.69 - 0.62    2108   3360     62.7     2.24      2.88    3.29  0.2890
0.62 - 0.57    1929   3542     54.5     1.44      1.51    1.79  0.4191
0.57 - 0.54    1123   2367     47.4     1.10      0.90    1.14  0.5593
----------------------------------------------------------------------
0.64 - 0.54    3720   7014     53.0     1.43      1.47    1.76  0.4170
 Inf - 0.54   12961  18171     71.3     5.08     20.64   13.61  0.0514

Merged [A],  lowest resolution = 11.49 Angstroms
```

CC1/2, and resolution cut-off

- CC1/2 should be close to 100% throughout resolution range
- where CC1/2 drops below 70%, noise becomes significant, and data at higher resolution can be excluded from refinement
- $I/\sigma(I)$ should be about 2, where CC1/2 about 70%
- $I/\sigma(I)$ should be about 1, where CC1/2 about 40% (in cases very resolution cut-off is critical)

Example $CC_{1/2}$, and resolution cut-off



CC1/2 *vs.* data resolution

R-values for the model

$$R = R1 = \sum_h \frac{||F_h(data)| - |F_h(model)||}{|F_h(data)|}$$

weighted R-value:

$$wR = \sum_h \frac{|w_h|F_h(data)| - |F_h(model)||}{w_h|F_h(data)|}$$

weighted intensity based R-value:

$$wR2 = R_B = \sqrt{\sum_h \frac{|w_h(I_h(data) - I_h(model))^2|}{w|I_h(data)|^2}}$$

Small molecules: $R1$ of the refined model 2-5 %.

GooF

Goodness of Fit

$$GooF = \sqrt{\frac{\sum_h w_h \left({F_h}^2(data) - {F_h}^2(model)\right)^2}{n-p}}$$

- Takes number of parameters (p) and number of data (n) into account
- Ideally ≈ 1, increases with worse model

model: residual density

SHELXL calculates the “highest peak” and “deepest hole” in the electron density map. Units are electrons, e.g. at the beginning of model building:

```
Electron density synthesis with coefficients Fo-Fc


Highest peak    4.95  at  0.5434  0.9981  0.3231  [  0.04 A from RU01
Deepest hole   -3.34  at  0.0057  0.4976  0.3299  [  0.99 A from RU02
^^^^^^
Mean =    0.00,   Rms deviation from mean =    0.34 e/A^3
^^^^^^^^^^^^
```

model: residual density

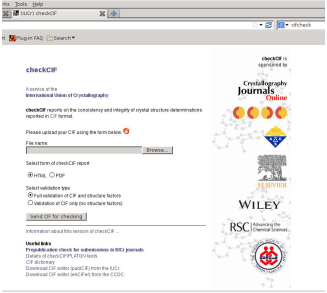
SHELXL calculates the “highest peak” and “deepest hole” in the electron density map. Units are electrons, e.g. for the refined model:

```
Electron density synthesis with coefficients Fo-Fc

Highest peak    0.50  at  0.6610  0.1969  0.4278  [  0.69 A from C006
Deepest hole   -1.22  at  0.2635  0.6156  0.2132  [  0.04 A from P003
^^^^^^


Mean =    0.00,   Rms deviation from mean =    0.06 e/A^3
^^^^^^^^^^^^
```

checkCIF (PLATON web-based)


rks Tools Help
(IUCr) checkCIF
cifcheck
rt Plug-in FAQ Search
Crystallography Journals Online
ELSEVIER
WILEY
RSC Advancing the Chemical Sciences
THE CHEMICAL SOCIETY OF JAPAN
NIPPON KAGAKUKAI
CSJ

checkCIF

A service of the
International Union of Crystallography

checkCIF reports on the consistency and integrity of crystal structure determinations reported in CIF format.

Please upload your CIF using the form below.

File name: Browse...

Select form of checkCIF report

● HTML ○ PDF

Select validation type

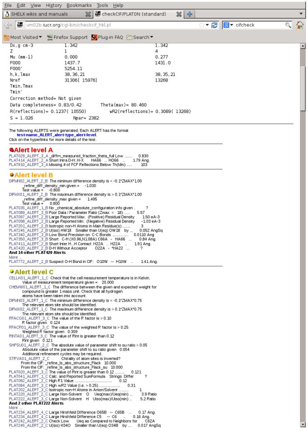
● Full validation of CIF and structure factors
○ Validation of CIF only (no structure factors)

Send CIF for checking

Information about this version of checkCIF ...

Useful links
Prepublication check for submissions to IUCr journals
Details of checkCIF/PLATON tests
CIF dictionary
Download CIF editor (publCIF) from the IUCr
Download CIF editor (enCIFer) from the CCDC

checkCIF is sponsored by


File Edit View History Bookmarks Tools Help
SHELX wikis and manuals
checkCIF/PLATON (standard)
vm02b.iucr.org/cgi-bin/checkcif_hkl.pl
cifcheck
Most Visited Firefox Support Plug-in FAQ Search

```
Dx,g cm-3            1.342                        1.342
Z                    1                            4
Mu (mm-1)            0.000                        0.277
F000                 1437.7                       1431.0
F000'                5254.11
h,k,lmax             38,36,21                     38,35,21
Nref                 31306[ 15976]                13268
Tmin,Tmax
Tmin'
Correction method= Not given
Data completeness= 0.83/0.42       Theta(max)= 80.460
R(reflections)= 0.1237( 10550)         wR2(reflections)= 0.3089( 13268)
S = 1.026                Npar= 2382
```

The following ALERTS were generated. Each ALERT has the format
test-name_ALERT_alert-type_alert-level.
Click on the hyperlinks for more details of the test.

Alert level A

```
PLAT029_ALERT_3_A _diffrn_measured_fraction_theta_full Low .......    0.830
PLAT414_ALERT_2_A Short Intra D-H..H-X      HA66   ..  HO66    ..     1.79 Ang.
PLAT910_ALERT_3_A Missing # of FCF Reflections Below Th(Min) .....     103
```

Alert level B

```
DIFMN02_ALERT_2_B  The minimum difference density is < -0.1*ZMAX*1.00
           _refine_diff_density_min given =     -1.030
           Test value =     -0.800
DIFMX01_ALERT_2_B  The maximum difference density is > 0.1*ZMAX*1.00
           _refine_diff_density_max given =      1.495
           Test value =      0.800
PLAT035_ALERT_1_B No _chemical_absolute_configuration info given .        ?
PLAT089_ALERT_3_B Poor Data / Parameter Ratio (Zmax  <   18) .....      5.57
PLAT097_ALERT_2_B Large Reported Max.  (Positive) Residual Density      1.50 eA-3
PLAT098_ALERT_2_B Large Reported Min.  (Negative) Residual Density     -1.03 eA-3
PLAT201_ALERT_2_B Isotropic non-H Atoms in Main Residue(s) .......         3
PLAT245_ALERT_2_B U(iso) HW18    Smaller than U(eq) OW18    by ...     0.052 AngSq
PLAT340_ALERT_3_B Low Bond Precision on  C-C Bonds ...............    0.0110 Ang
PLAT350_ALERT_3_B Short   C-H (X0.96,N1.08A) C66A   -   HA66   ...      0.84 Ang.
PLAT411_ALERT_2_B Short Inter H...H Contact  H22A   ..  H22A    ..     1.91 Ang.
PLAT420_ALERT_2_B D-H Without Acceptor       O22A   -  *HA22   ....        ?
```

And 14 other PLAT420 Alerts
More ...

```
PLAT772_ALERT_2_B Suspect O-H Bond in CIF:   O10W   --  H10W    ..      1.41 Ang.
```

Alert level C

```
CELLK01_ALERT_1_C  Check that the cell measurement temperature is in Kelvin.
           Value of measurement temperature given =    20.000
CHEMW01_ALERT_1_C  The difference between the given and expected weight for
           compound is greater 1 mass unit. Check that all hydrogen
           atoms have been taken into account.
DIFMN03_ALERT_1_C  The minimum difference density is < -0.1*ZMAX*0.75
           The relevant atom site should be identified.
DIFMX02_ALERT_1_C  The maximum difference density is > 0.1*ZMAX*0.75
           The relevant atom site should be identified.
RFACG01_ALERT_3_C  The value of the R factor is > 0.10
           R factor given   0.124
RFACR01_ALERT_3_C  The value of the weighted R factor is > 0.25
           Weighted R factor given   0.309
RINTA01_ALERT_3_C  The value of Rint is greater than 0.12
           Rint given   0.121
SHFSU01_ALERT_2_C  The absolute value of parameter shift to su ratio > 0.05
           Absolute value of the parameter shift to su ratio given   0.054
           Additional refinement cycles may be required.
STRVA01_ALERT_2_C          Chirality of atom sites is inverted?
          From the CIF: _refine_ls_abs_structure_Flack    10.000
          From the CIF: _refine_ls_abs_structure_Flack_su   10.000
PLAT020_ALERT_3_C The value of Rint is greater than 0.12 .........      0.121
PLAT041_ALERT_1_C Calc. and Reported SumFormula    Strings  Differ         ?
PLAT082_ALERT_2_C High R1 Value ..................................      0.12
PLAT084_ALERT_3_C High wR2 Value (i.e. > 0.25) ...................      0.31
PLAT202_ALERT_3_C Isotropic non-H Atoms in Anion/Solvent .........         1
PLAT220_ALERT_2_C Large Non-Solvent    O     Ueq(max)/Ueq(min) ...       3.9 Ratio
PLAT222_ALERT_3_C Large Non-Solvent    H    Uiso(max)/Uiso(min) ..       5.2 Ratio
```

And 2 other PLAT222 Alerts
More ...

```
PLAT234_ALERT_4_C Large Hirshfeld Difference O65B   --  C65B    ..      0.17 Ang.
PLAT234_ALERT_4_C Large Hirshfeld Difference C5     --  C6      ..      0.16 Ang.
PLAT242_ALERT_2_C Check Low       Ueq as Compared to Neighbors for       C62A
PLAT245_ALERT_2_C U(iso) H34O    Smaller than U(eq) O34B    by ...     0.017 AngSq
```

Every published structure *should* have a checkCIF report. There are different alert levels of decreasing severity. Reviewers typically require that a structure should **not** contain A- or B-alerts.

Summary

- A model without data does not reflect data quality
- Data quality: data resolution, multiplicity, R-values, I/σ_I, $\mathrm{CC}_{1/2}$
- Model quality: R1-values, GooF, residual density
- available for everyone: checkCIF `http://checkcif.iucr.org` (with or without data)
- *ALERT levels* A, B, ...
- (Analogously for macromolecular structures: `http://molprobity.biochem.duke.edu/`)

End of lecture