Chemical Crystallography and Structural Chemistry

VO 270063-1

Lecture № 6 — 4th May 2023

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

tim.gruene@univie.ac.at

Course Schedule

2nd	March	Lecture № 1	9th	March	Lecture № 2
16th	March	Lecture № 3	23th	March	Exercise № 1
30st	March	Lecture № 4	20th	April	Lecture № 5
27th	April	Exercise № 2	4th	May	Lecture № 6
11th	May	Lecture № 7	25th	May	Exercise № 3
1st	June	**no lecture**	15th	June	Lecture № 8
22nd	June	Lecture № 9	29th	June	Exercise № 4

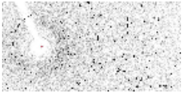
Previous Lecture

- Scaling: Idealisation of the experiment
- Phasing: initial chemical model
- Phasing: Patterson map

Contents

1 Structure Refinement

Data collection → **Data integration** → **Data Scaling**



several GB

```
0 0 -1      2.7  0.9
0 0  1      4.0  1.0
0 0 -2 1'257.0 35.5
0 0 -2 1'600.0 42.7
```

several files, 100’s MB

```
0 0 -1      2.8 0.55
0 0  1      3.8 0.63
0 0 -2 1'432.0 95.7
0 0 -2 1'282.0 85.9
```

1 “hkl”-file, 50MB

→ **Phasing**

Starting model

→ **Refinement**

Chemically sensible model

Model Building & Refinement

- Refinement optimises computationally the parameters of the structure with respect to the data
- Model building make modifications that are too large for computer optimisation, e.g.
  1. Addition or removal of atoms
  2. correction of atom types
  3. modelling of disorder and multiple conformations

Structure parameters

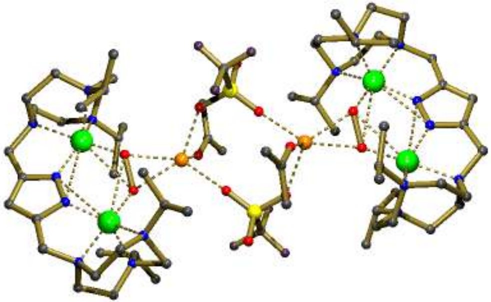
A "structure" consists of a set of parameters, *i.e.* numbers. Refinement improves these numbers for make the structure better correspond to the data.

```
X       Y       Z   occ.
N2A  2 0.8142  0.9066  0.8201 11.00000 =
0.0497  0.0413  0.0363 -0.0136  -0.0041 -0.0063
U11     U22     U33     U23      U13     U12
```

1. 3 atom coordinates x, y, z
2. 6 atomic displacement parameters ADP $\begin{pmatrix} U_{11} & U_{12} & U_{13} \\ & U_{22} & U_{23} \\ & & U_{33} \end{pmatrix}$
3. possibly 1 occupancy parameter for disorder

9 Parameters per atom of the asymmetric unit are being refined, plus extra parameters in case of disorder or other special circumstances

Example for parameters [1]



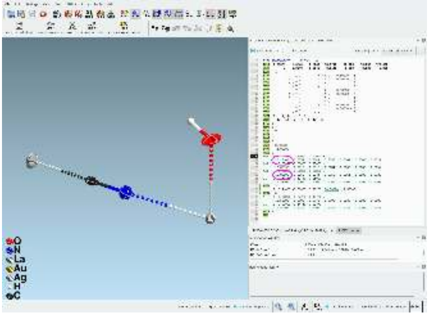
- $C_{34}H_{63}Cu_2F_6N_8NaO_{9.5}S_2$
- 62.5 non-hydrogen atoms
- 724 parameters
- hydrogen atoms are “special”

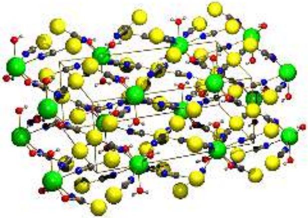
Atom occupancy — symmetry

- Asymmetric unit: average of all asymmetric units of the crystal
- Molecules do not always strictly follow symmetry
- Some atoms sit on "special position", i.e. fix points of symmetry elements. Their occupancy is divided by the multiplicity of the symmetry element
- e.g. atom on three-fold axis: occupancy 33 %

Atom occupancy — alternative elements

Example: $La[Ag_{0.39}Au_{0.61}(CN)_2]_3 \cdot H_2O$ [2] 39% of all unit cells contain Ag, 61% contain Au at the same position

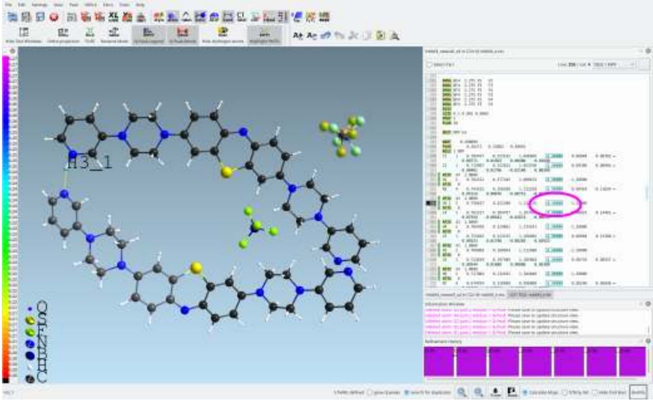




occ(Au1) + occ(Ag1) = 0.154 + 0.096 = 1/4 with 4-fold multiplicity

Atom occupancy

Example: Disordered BF_4^- and one H-atom on special position [3]



Refinement = improvement of parameters

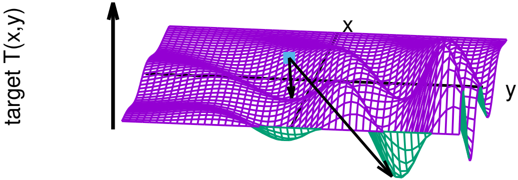
Computationally, refinement minimises the discrepancy between the observed data $\mathrm{I}_{\mathrm{obs}}$ and the calculated data $\mathrm{I}_{\mathrm{calc}}$. $\mathrm{I}_{\mathrm{calc}}$ is calculated from the model parameters, mainly atom coordinates x, y, z and atomic dispersion parameters ADPs

$$T(\vec{x}_i, U_i, (\text{occupancies}, \ldots)) = \sum_{(hkl)} w(hkl)|I_{\mathsf{obs}}(hkl) - I_{\mathsf{calc}}(hkl)|^2$$

$w(hkl)$ downweights untrusted reflections, typically $w(hkl) = 1/\sigma_I(hkl)$.
Note: different refinement programs use different target functions.

Least-square-minimisation

The shape of the target function $T = \sum_{(hkl)} w(hkl)|I_{\text{obs}}(hkl) - I_{\text{calc}}(hkl)|^2$ enables optimisation based on least-squares method (`L.S.` command in SHELXL).

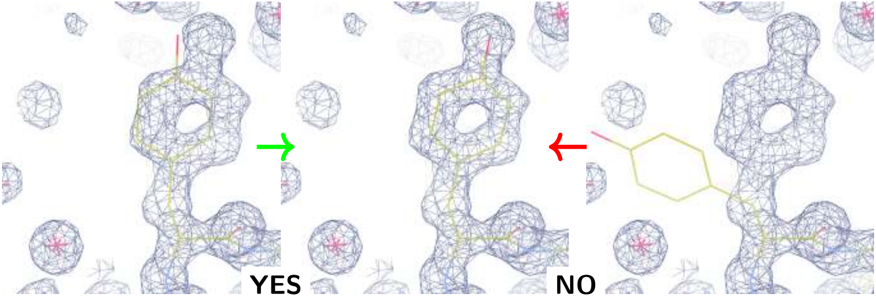


The algorithm finds the next minimum, but cannot jump across humps.

Advantages and limitations of refinement

- Refinement finds the “next” local minimum
- only small changes in the structure
- does not add or remove atoms
- no change of element types
- one never knows whether the optimum is reached. However, for small molecules, the starting model usually converges to a good model.

The “next” local minimum



Model building

- manual modifications “help” refinement cross local humps
    - large movements of individual atoms (out of local traps)
    - delete wrong atoms
    - add missing atoms
    - correct atom type
- model building = add chemical understanding to the model
- graphic programs (Olex2, shelXle, Crystals, JANA2020, Crystls …[4, 5, 6, 7])
- guided by the electron density map

Iterative process: improve model -> refine -> improve model -> refine -> …

Electron density map and difference map

$$\rho(x,y,z) = FT(|F_{\text{obs}}(hkl)|, \phi_{\text{calc model}}(hkl))$$

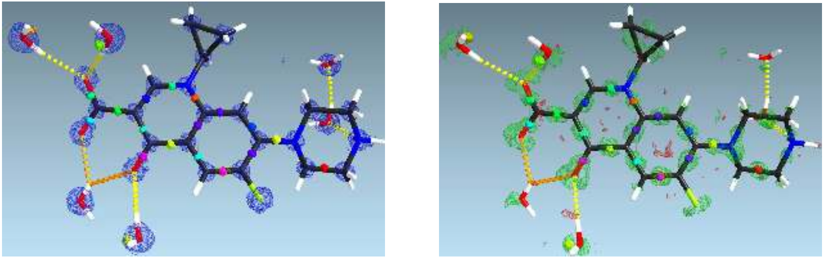
Fourier transformation from measured structure factor amplitudes $|F_{\text{obs}}(hkl)|$ and calculated phases $\phi_{\text{calc model}}(hkl)$
This model should follow this map.
The map

$$\Delta\rho(x,y,z) = FT(|F_{\text{obs}}(hkl)| - |F_{\text{calc}}(hkl)|, \phi_{\text{calc}}(hkl))$$

is called **difference map**. It reveals discrepancies between the model and the data.
Model building and refinement aim at reducing these discrepancies.

Example map: Ciprofloxacin

Structure of Ciprofloxacin, [8], ultra high resolution 0.43 Å



$\rho(x,y,z)$ (usually blue mesh) $\Delta\rho(x,y,z)$ (usually green / red mesh)

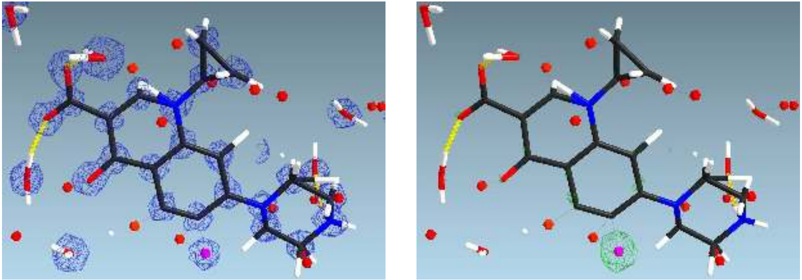
positive $\Delta\rho$: Model misses something. SHELXL places **Q-peaks**

negative $\Delta\rho$: model contains too much

Example map: Ciprofloxacin

Structure of Ciprofloxacin, [8], ultra high resolution 0.43 Å

- data resolution truncated to 0.9 Å
- Fluorine atom F removed from model

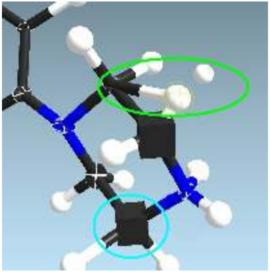


$\rho(x, y, z)$ (blue mesh) $\Delta\rho(x, y, z)$ (red / green mesh)

Refinement without restraints

$$T(\vec{x}_i, \mathbf{U}_i) = \sum_{(hkl)} w(hkl)(I_{\text{data}}(hkl) - I_{\text{model}}(hkl))^2$$

This formula carries out **unrestrained refinement**, purely taking experimental data into account. With poor data, this **can** cause



- unrealistic bond distances and bond angles
- negative ADPs (cubes) are physically meaningless
- refinement can produce non-sense results

Unrestrained refinement, example

Unrestrained refinement of protein structure with 1.4 Å resolution

Data to parameter ratio

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

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

0.43 Å resolution 26'308 reflections. $26'308 : 540 = 48.7$ data points per parameter: very high data to parameter ratio, data sufficient to produce chemically sensible structure

0.8 Å resolution 2'926 reflections. $2'926 : 540 = 5.4$ data points per parameter: low data to parameter ratio, data insufficient to produce chemically sensible structure

Chemically sensible part needs to be **restrained** -> restrained refinement

Restrained refinement

Except for at very high resolution, the refinement program has to be told some chemistry to make sure the model remains chemically meaningful. There are two different types how this can be accomplished:

Constraints Express an equality and permit no deviation from fixed value

Restraints Express similarity and provide some flexibility from target value.

Restraints are much more common than constraints

Constraints

- The structure of $La[Ag_{0.39}Au_{0.61}(CN)_2]_3{\cdot}H_2O$ has either gold or silver at one location.
- In every unit cell there is always one atom at this location

$$\begin{aligned} occ(Au) + occ(Ag) &= 1 \\ occ(Au) &= 1 - occ(Ag) \end{aligned}$$

- Only the occupancy of silver has to be determined. The occupancy of gold can be calculated (or *vice versa*)
- remark: the program SHELXL uses the command FVAR (“free variables”) to realise constraints.

Each constraint reduces the number of parameters by 1

Important constraints

negative ADP value, mainly for hydrogen atoms: U(HA)= 1.2*U(CA)

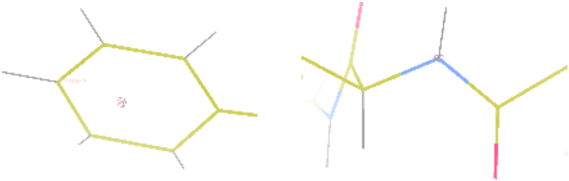
```
CA     1     0.673087     0.878303     0.111632     11.00000     0.31129
HA     6     0.679625     0.855075     0.095775     11.00000    -1.20000
```

hydrogen positions: `AFIX`

```
N      3     0.611916     1.012005     0.052456     11.00000     0.18165
AFIX   43
H      6     0.628491     1.011598     0.033498     11.00000    -1.20000
AFIX    0
CB     1     0.622779     1.076653     0.067974     11.00000     0.18216
AFIX   23
HB1    6     0.608063     1.103479     0.072220     11.00000    -1.20000
HB2    6     0.641195     1.080130     0.047994     11.00000    -1.20000
AFIX    0
```

AFIX: riding atom model

- Except for at very high resolution ($d \ll 0.8$ Å), hydrogen atoms are invisible to X-rays
- the positions of most hydrogen atoms can be calculated: bond distances are known from spectroscopy, positions are determined by reducing steric clashes
- Advantages: hydrogen atoms do not add parameters, the contribute to VdW repulsion (BUMP command), they have a small, but non-zero contribution to the scattering.



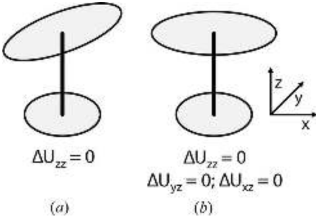
Restraints: Geometry

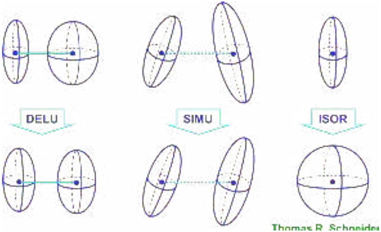
- restraints can be expressed as inequality "$\leq$"
- best known restraints: R. A. Engh, R. Huber, *Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement*, Acta Crystallogr. (1991), **A**47, pp. 392–400; *e.g.*

$$|d(N, C_\alpha) - 1.458\text{Å}| \leq 0.02 \qquad |d(C_\alpha, C_\beta) - 1.521\text{Å}| \leq 0.02$$

Restraints: ADP values [9]

- restraints for ADPs: chemical bond affects thermal vibrations





DELU, SIMU, ISOR (Thomas R. Schneider)

Restraints resemble data

Restraints are treated with additional terms to the target function:

$$T(\vec{x}_i, \mathbf{U}_i) = \sum_{hkl} w_{hkl}(I_{\text{data}}(hkl) - I_{\text{model}}(hkl))^2 + W \sum_{\text{N.B. } i} w_i(T_i^{\text{data}} - \langle T_i \rangle)^2$$

Restraints act like additional data points

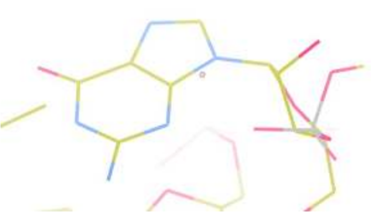
- W weights restraints and observed data
- the higher the resolution, the lower weight W
- the expected mean values $\langle T_i \rangle$ can be derived statistically from high resolution structures, or sometimes can be computed quantum chemically

Summary refinement & model building

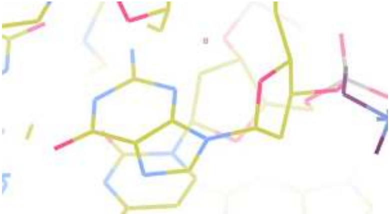
- model building improves the model in large steps
- refinement optimises the model against the data
- constraints and restraints are used to ensure a chemically reasonable model
- constraints reduce the number of parameters, restraints act like data: both increase the data to parameter ratio

2 Model quality and data quality: structure validation

Atom coordinates $\neq$ model accuracy



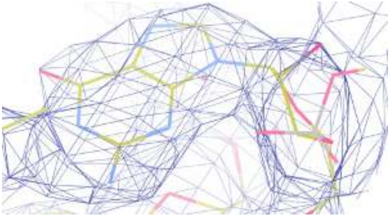
Guanine model in ribosome, data resolution 3.1Å



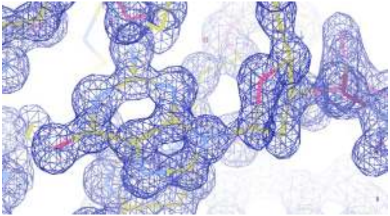
Guanine model in Z-DNA, at resolution 1.0 Å

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, $\alpha = \beta = \gamma = 90°$, $P4_12_12$)

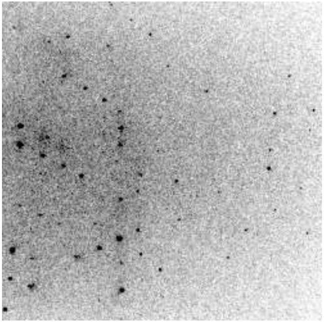
- PDB ID 1J5E: 51'atoms atoms $=$ 207'768 parameters
- data resolution 3.05 Å 238'205 reflections

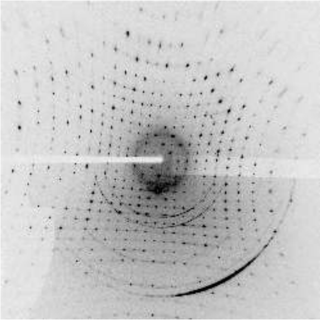
$$\frac{238'205}{207'768} = 1.15$$

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

3 Indicators for data quality

Example data quality





Important quality indicators

R$_{\text{meas}}$ relative difference between symmetry equivalent reflections and their mean value

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 $\mathrm{R_{sym}}$:

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

$\mathrm{R_{int}}$ 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:

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

Example data statistics (`XPREP`)

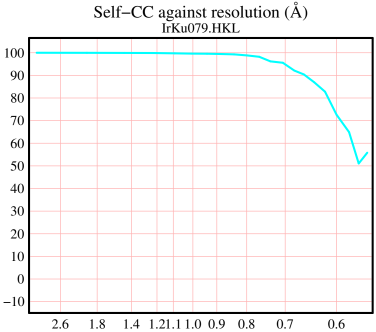
```
Resolution  #Data #Theory %Complete Redundancy Mean I Mean I/s Rmerge
Inf - 2.46     196    197     99.5    39.27    215.01  110.27  0.0300
2.46 - 1.13    1762   1825     96.5    14.86     75.32   42.01  0.0453
1.13 - 0.89    1972   2123     92.9     8.71     25.52   19.00  0.0895
0.89 - 0.77    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

A good quality crystal diffracts beyond the theoretical limit $d_{\text{min}} = \lambda/2$. Resolution cut-off is not an issue, one can use all data. Large complexes, supramolecular structures, low quality crystals reach the diffraction limit before the theoretical limit. One has to decide where to cut the diffraction data.

- 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)

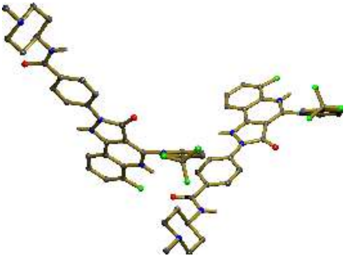
Example $CC_{1/2}$, and resolution cut-off

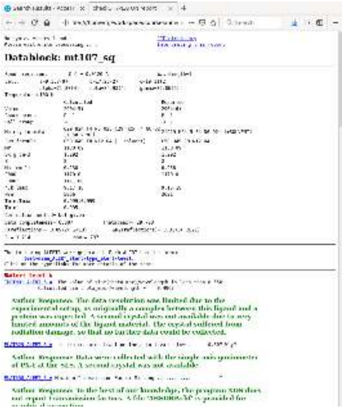


CC1/2 *vs.* data resolution; plot generated with XPREP

4 Indicators for model quality

Model quality [10]





Datablock: mt107_sq

Author Response: The data resolution was limited due to the experimental setup, as originally a complex between this ligand and a protein was expected. A second crystal was not available due to very limited amounts of the ligand material. The crystal suffered from radiation damage, so that no further data could be collected.

Author Response: Data were collected with the single axis goniometer at PX-I at the SLS. A second crystal was not available.

Author Response: To the best of our knowledge, the program XDS does not report transmission factors.

R-values for the model

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

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 %.

supramolecules compounds, MOFs, ... $R1$ of the refined model **can** be highter, 2-15 %

macromolecular compounds $R1$ of the refined model 15-25 %

To a great extent, this discrepancy is due to the unmodelled solvent region in the latter two types of compounds

Goodness of Fit — GooF

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

- 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 `https://checkcif.iucr.org/`

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 Validation

- A model without data does not reflect data quality
- Data quality: data resolution, multiplicity, R-values, I/σ_I, 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/`)

References

[1] K. Dalle et al. 'A weakly coupled biologically relevant $Cu_2^{II}(\mu-\eta^1:\eta^1-O_2)$ *cis*-peroxo adduct that binds side-on to additional metal ions'. In: *J. Am. Chem. Soc.* 136 (2014), pp. 7428–7434. DOI: `10.1021/ja5025047`.

[2] Julie Clarissa F. Colis et al. 'Structural studies of lanthanide ion complexes of pure gold, pure silver and mixed metal (gold–silver) dicyanides'. In: *Dalton Trans.* (2005), pp. 675–679. DOI: `10.1039/B413967D`.

[3] Tim Gruene et al. 'Rapid structure determination of microcrystalline molecular compounds using electron diffraction'. In: *Angew. Chem., Int. Ed.* 57 (2018), pp. 16313–16317. DOI: `10.1002/anie.201811318`.

[4] *Olex2*. Version 1.5. URL: `https://www.olexsys.org/olex2` (visited on 07/06/2022).

[5] C. Hübschle. *ShelXle*. URL: `https://www.shelxle.org` (visited on 07/06/2022).

[6] Václav Petříček, Michal Dušek and Lukáš Palatinus. *JANA2020*. URL: `http://jana.fzu.cz` (visited on 07/06/2022).

[7] Paul W. Betteridge et al. *CRYSTALS*. URL: `http://www.xtl.ox.ac.uk/crystals.html` (visited on 07/06/2022).

[8] Julian Jacob Holstein, Christian Bertram Hübschle and Birger Dittrich. ‘Electrostatic properties of nine fluoroquinolone antibiotics derived directly from their crystal structure refinements’. In: *CrystEngComm* 14 (2012), pp. 2520–2531. DOI: `10.1039/C1CE05966A`.

[9] Andrea Thorn, Birger Dittrich and George M. Sheldrick. ‘Enhanced rigid-bond restraints’. In: *Acta Crystallogr* A68 (2012), pp. 448–451. DOI: `10.1107/S0108767312014535`.

[10] Marco F. Taddio et al. ‘Synthesis and Structure–Affinity Relationship of Small Molecules for Imaging Human CD80 by Positron Emission Tomography’. In: *J. Med. Chem.* 62 (2019), pp. 8090–8100. DOI: `10.1021/acs.jmedchem.9b00858`.