Chemical Crystallography and Structural Chemistry

VO 270287

Lecture № 8 — 23rd June 2022

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

tim.gruene@univie.ac.at

Course Details

3rd	March	Lecture № 1	10th	March	Lecture № 2
17th	March	**no lecture**	24th	March	Exercise № 1
31st	March	Lecture № 3			
7th	April	Discussion Ex. 1	14th	April	**Easter break**
21st	April	**Easter break**	28th	April	Lecture № 4
5th	May	Power cut	12th	May	**no lecture**
19th	May	Lecture № 6	26th	May	**Ascension Day**
2nd	June	Exercise № 2	9th	June	Lecture № 7
16th	June	**Corpus Christi**	23th	June	Lecture № 8
30th	June	**no lecture**			

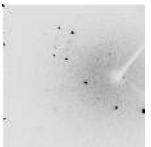
Previous Lecture

- Phasing methods: Patterson map and direct phasing
- Structure refinement - mathematical concepts
- Model building

Contents

1 Structure Refinement

Data collection ➡ Data integration ➡ Data Scaling



```
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
```

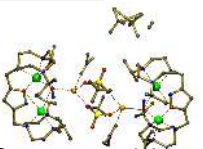
```
0 0 -1      2.8 0.55
0 0  1      3.8 0.63
0 0 -2 1'432.0 95.7
0 0 -2 1'282.0 85.9
```

several GB

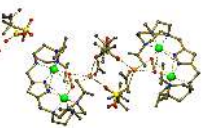
several files, 100's MB

1 “hkl”-file, 50MB

➡ Phasing ➡ **Refinement**



Starting model



Chemically sensible model

Electron density map and difference map

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

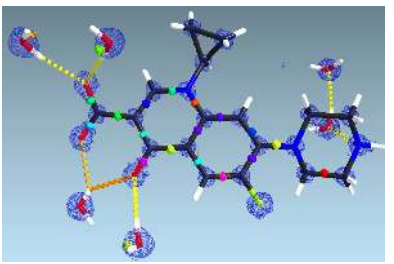
Fourier transformation from measured structure factor amplitudes $|F_{\text{obs}}(hkl)|$ and calculated phases $\phi_{\text{calc model}}(hkl)$
This model should follow this map.
The map

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

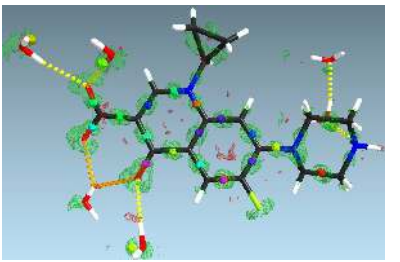
is called **difference map**. It reveals discrepancies between the model and the data.
Model building and refinement aim at reducing these discrepancies.

Example map: Ciprofloxacin

Structure of Ciprofloxacin, [1], ultra high resolution 0.43 Å



$\rho(x, y, z)$ (usually blue mesh)



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

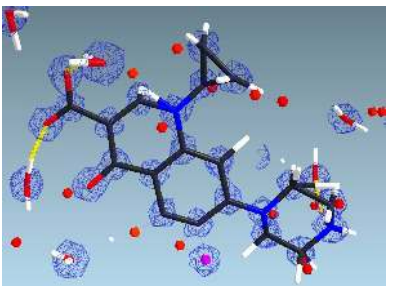
positive $\Delta\rho$: Model misses something. SHELXL places **Q-peaks**

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

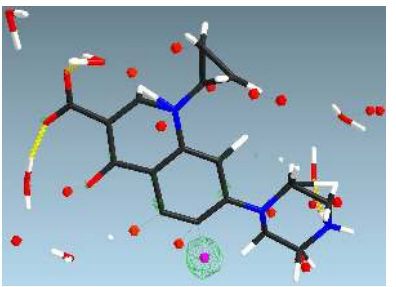
Example map: Ciprofloxacin

Structure of Ciprofloxacin, [1], 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)

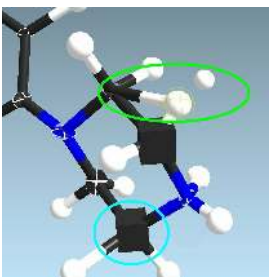


$\Delta\rho(x, y, z)$ (red / green mesh)

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

$$occ(Au) + occ(Ag) = 1$$
$$occ(Au) = 1 - occ(Ag)$$

- 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

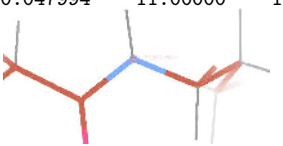
Important constraints

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.

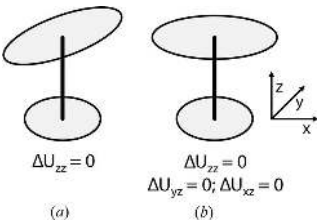
Restraints: Geometry

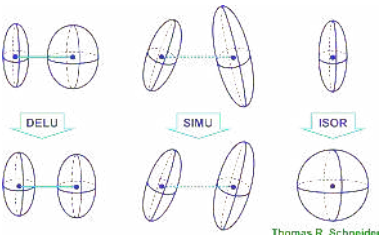
- restraints can be expressed as inequality "$\leq$"
- best known restraints: R. A. Engh, R. Huber, *Accurate Bond and Angle Parameters for X-ray Protein Structure Refinement*, Acta Crystallogr. (1991), **A**47, pp. 392–400; *e.g.*

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

Restraints: ADP values [2]

- 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

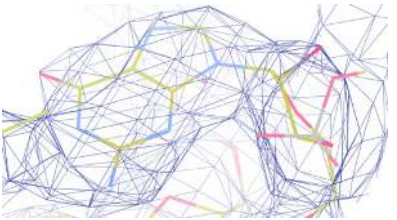
Atom coordinates $\neq$ model accuracy

Guanine model in ribosome, data resolution 3.1Å

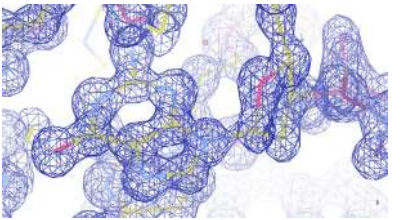
Guanine model in Z-DNA, at resolution 1.0 Å

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

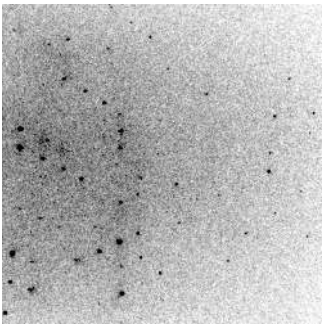
- PDB ID 1J5E: 51'atoms atoms = 207'768 parameters
- data resolution 3.05 Å 238'205 reflections

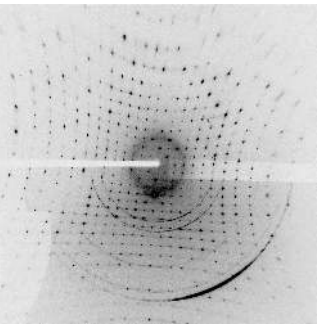
$$\frac{238'205}{207'768} = 1.15$$

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_{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

These values are typically listed for the overall data set as well as broken down into resolution shells.

R_{meas} (Diederichs and Karplus [3])

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

Symmetry equivalent should have identical intensities. R_{meas} indicates the level of deviation from this assumption. Note that dynamical diffraction (see below) breaks this assumption.

Some crystallographers, in particular in chemical crystallography, R_{int}, alias R_{merge} or R_{sym} is cited instead.

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

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

Example data statistics (XPREP- LHP281)

Resolution	#Data	#Theory	%Complete	Redundancy	Mean I	Mean I/s	Rmerge
Inf - 3.20	182	184	98.9	17.32	1489.67	276.27	0.0198
3.20 - 1.48	1656	1656	100.0	8.96	668.40	74.08	0.0228
1.48 - 1.17	1872	1872	100.0	6.01	344.26	25.05	0.0433
1.17 - 1.03	1705	1716	99.4	5.02	196.98	11.91	0.0710
1.03 - 0.93	1891	1934	97.8	4.40	146.74	7.49	0.0942
0.93 - 0.86	1878	1956	96.0	3.88	109.30	4.75	0.1262
0.86 - 0.77	2849	3461	82.3	2.91	67.22	2.64	0.1811
0.87 - 0.77	3120	3748	83.2	2.97	69.75	2.77	0.1759
Inf - 0.77	12033	12779	94.2	5.01	252.02	22.50	0.0411

Merged [A], lowest resolution = 11.35 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
- in desperate cases, data can be cut where CC1/2 drops below 40%

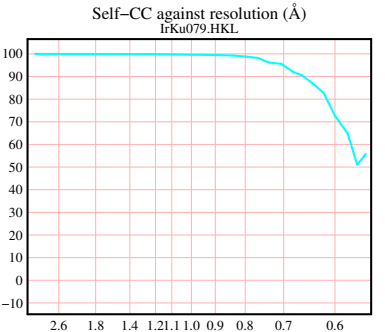
I/σ and CC1/2 (Karplus and Diederichs [4])

- CC1/2 determines the cross-correlation coefficient between equivalent reflections
- it is independent of the values of σ
- CC1/2 can be used to estimate the resolution of the data set: where CC1/2<40 %, mainly noise has been integrated
- CC1/2 also measures the quality of scaling:

$$CC1/2 \approx 70\% \Leftrightarrow I/\sigma \approx 2$$
$$CC1/2 \approx 40\% \Leftrightarrow I/\sigma \approx 1$$

deviation indicate improper σ values, i.e. the scaling model should be improved.

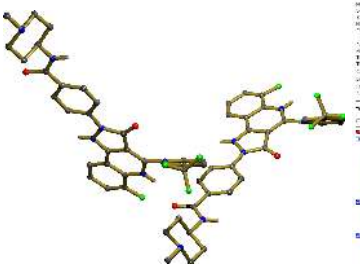
Example $CC_{1/2}$, and resolution cut-off



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

4 Indicators for model quality

Model quality [5]



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

Residual density — before

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
                                              ^^^^^^^^^^^^
```

Residual density — after

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/`)

5 Overview of additional topics in crystallography

Anomalous dispersion and chirality	Ariëns [6] and Spek [7]
Twinning	Sevvana et al. [8] and Nespolo, Ferraris and Souvignier [9]
Polymorphism, crystal engineering	Bernstein [10], Desiraju [11] and Hilfiker and Raumer [12]
Incommensurate crystals and quasicrystals	Janssen, Chapuis and Boissieu [13] and Steurer and Deloudi [14]
High-pressure crystallography	Katrusiak [15]
Quantum crystallography (charge density refinement)	Grabowsky, Genoni and Bürgi [16]
Neutron Crystallography	Blakeley [17]
Electron Crystallography	Gemmi et al. [18], Gruene et al. [19] and Gruene and Mugnaioli [20]

References

[1] 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.

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

[3] K. Diederichs and A. Karplus. 'Improved *R*–factors for diffraction data analysis in macromolecular crystallography'. In: *Nat. Struct. Biol.* 4 (1997), pp. 269–275.

[4] P. Andrew Karplus and Kay Diederichs. 'Linking Crystallographic Model and Data Quality'. In: *Science* 336 (2012), pp. 1030–1033.

[5] 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.

[6] E. J. Ariëns. 'Stereochemistry, a basis for sophisticated nonsense in pharmacokinetics and clinical pharmacology'. In: *Eur. J. Clin. Pharmacol.* 26 (1984), pp. 663–668.

[7] A. L. Spek. 'Absolute structure determination: pushing the limits'. In: *Acta Crystallogr.* B72 (2016), pp. 659–660. DOI: 10.1107/S2052520616014773.

[8] Madhumati Sevvana et al. 'Non-merohedral twinning: from minerals to proteins'. In: *Acta Crystallogr.* D75 (2019), pp. 1040–1050. DOI: 10.1107/S2059798319010179.

[9] Massimo Nespolo, Giovanni Ferraris and Bernd Souvignier. 'Effects of merohedric twinning on the diffraction pattern'. In: *Acta Crystallogr.* A70 (2014), pp. 106–125. DOI: 10.1107/S2053273313029082.

[10] J. Bernstein. *Polymorphism in Molecular Crystals*. Oxford University Press, 2007. ISBN: 9780199236565.

[11] Gautam R. Desiraju. 'Cryptic crystallography'. In: *Nat. Mater.* 1 (2002), pp. 77–79. DOI: 10.1038/nmat726.

[12] R. Hilfiker and M. v. Raumer, eds. *Polymorphism in the Pharmaceutical Industry. Solid Form and Drug Development*. Wiley-VCH, 2019. ISBN: 978-3-527-34040-8.

[13] T. Janssen, G. Chapuis and M. de Boissieu. *Aperiodic crystals*. Oxford University Press, 2018. ISBN: 9780198824442.

[14] Walter Steurer and Sofia Deloudi. 'Fascinating quasicrystals'. In: *Acta Crystallogr.* A64 (2008), pp. 1–11. DOI: 10.1107/S0108767307038627.

[15] Andrzej Katrusiak. 'High-pressure crystallography'. In: *Acta Crystallogr.* A64 (2008), pp. 135–148. DOI: 10.1107/S0108767307061181.

[16] Simon Grabowsky, Alessandro Genoni and Hans-Beat Bürgi. 'Quantum crystallography'. In: *Chem. Sci.* 8 (2017), pp. 4159–4176. DOI: 10.1039/C6SC05504D.

[17] M. P. Blakeley. 'Neutron macromolecular crystallography'. In: *Crystallogr. Rev.* 15 (2009), pp. 157–218. DOI: 10.1080/08893110902965003.

[18] M. Gemmi et al. '3D Electron Diffraction: The Nanocrystallography Revolution'. In: *ACS Cent. Sci.* 5 (2019), pp. 1315–1329.

[19] T. Gruene et al. 'Establishing electron diffraction in chemical crystallography'. In: *Nat. Rev. Chem.* 5 (2021), pp. 660–668. DOI: 10.1038/s41570-021-00302-4.

[20] Tim Gruene and Enrico Mugnaioli. '3D Electron Diffraction for Chemical Analysis: Instrumentation Developments and Innovative Applications'. In: *Chem. Rev. (Washington, DC, U. S.)* 121 (2021), pp. 11823–11834. DOI: 10.1021/acs.chemrev.1c00207.