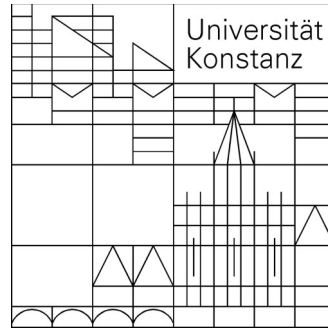


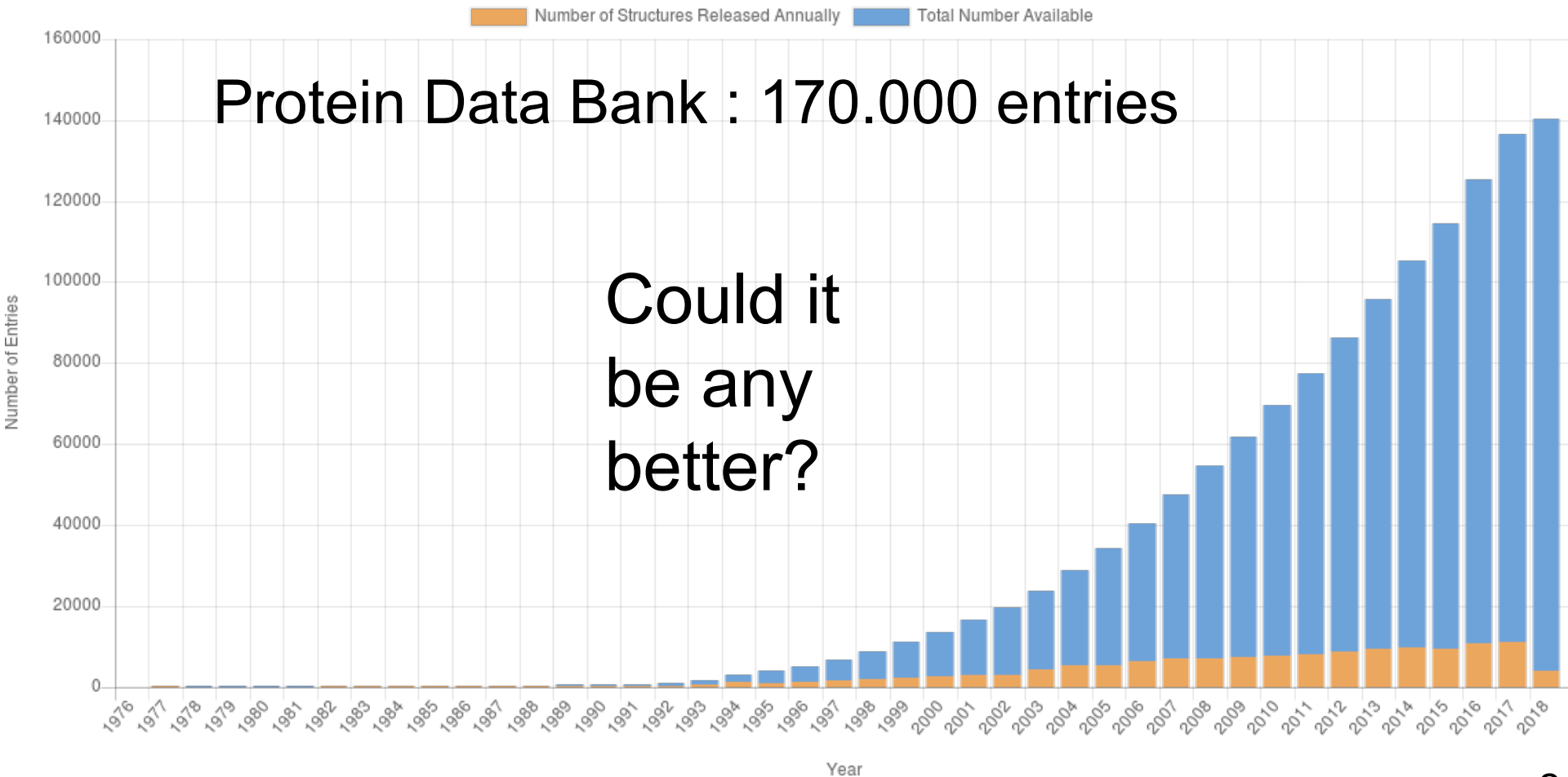
Assessing data quality – noise, errors and mistakes

Kay Diederichs



Protein Crystallography /
Molecular Bioinformatics
University of Konstanz, Germany

Crystallography has been extremely successful

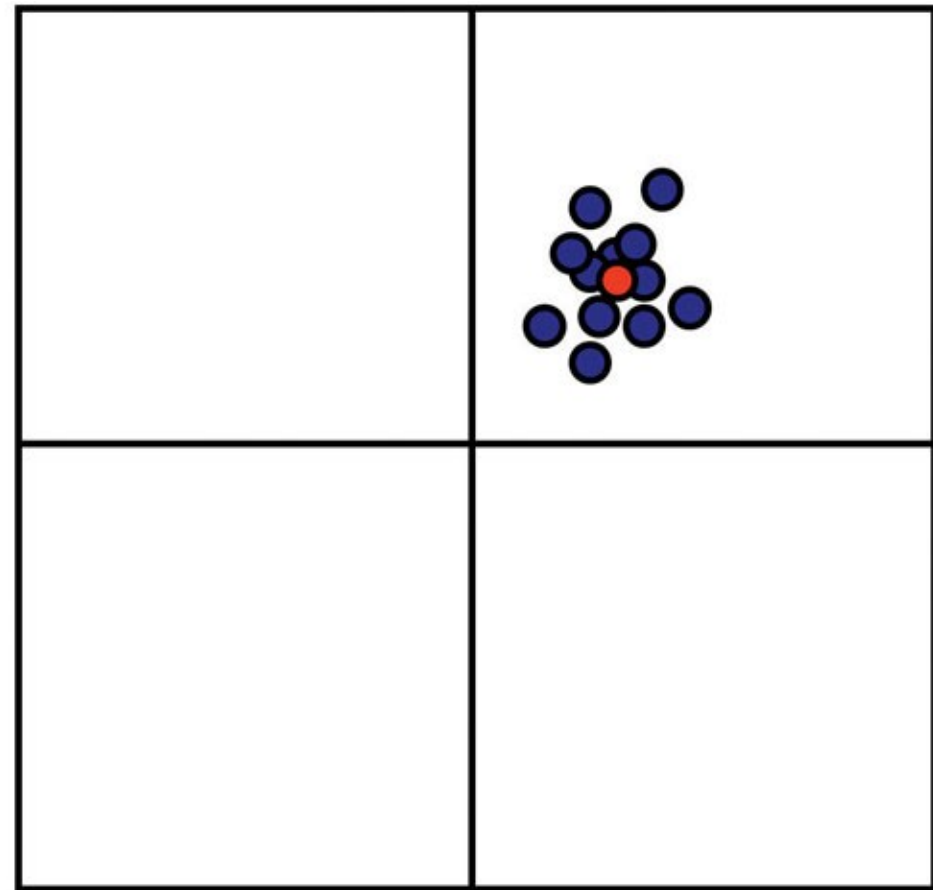
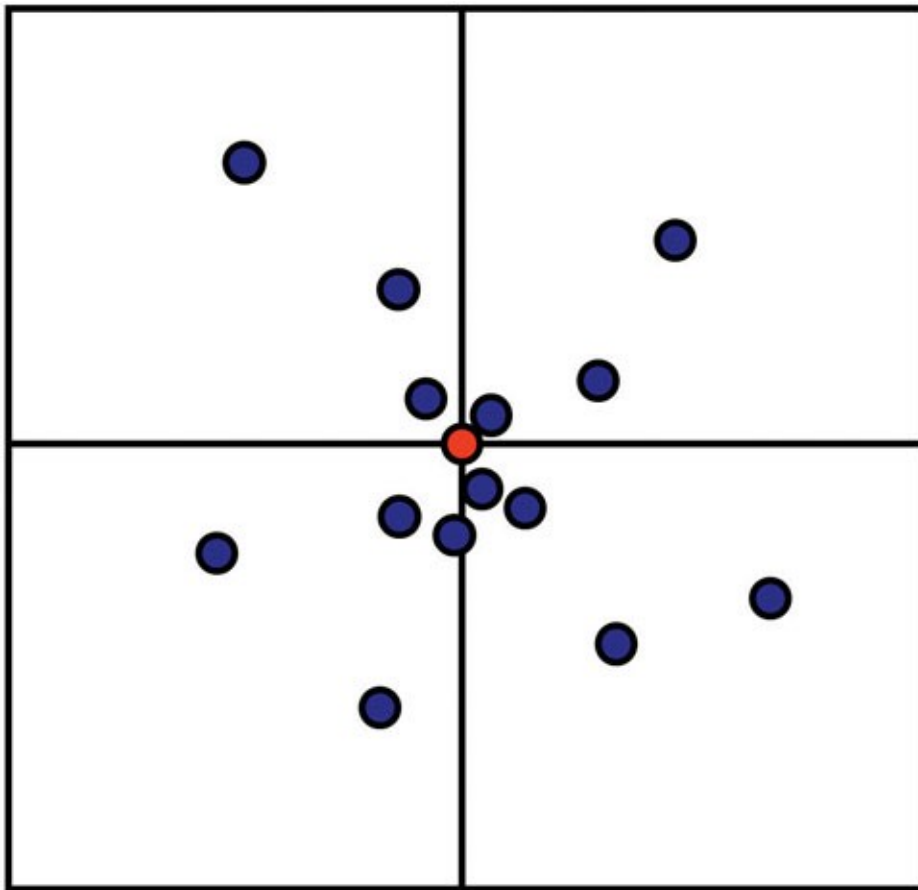


Four examples for

- *Rules* that may have been useful in the past under different circumstances, but are still commonly used today and result in wrong decisions
- *Concepts* resulting from first principles that would, if applied, deliver the information to reach the correct decision

1st example: Not understanding the difference between, and the relevance of **precision** and **accuracy**

“Quality”



Accuracy
Precision

– how different from the *true value*?
– how different are *measurements*?

Numerical example

Repeatedly determine $\pi=3.14...$ as 3.1, 3.2, 3.0 :

observations have **medium precision, medium accuracy**

Precision= relative |deviation| from average value=
 $(0+0.1+0.1)/(3.1+3.2+3.0) = 2.2\%$

Accuracy= average relative |deviation| from true value:
 $=1/3*(|3.14-3.1| + |3.14-3.2| + |3.14-3.0|)/3.14 = 2.5\%$

R_{merge}
formula!

$$R_{\text{merge}} = \frac{\sum_{hkl} \sum_{i=1}^n |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^n I_i(hkl)}$$

Repeatedly determine $\pi=3.14...$ as 2.70, 2.71, 2.72 :

observations have **high precision, low accuracy.**

Precision= relative |deviation| from average value=
 $(0.01+0+0.01)/(2.70+2.71+2.72) = 0.24\%$

Accuracy= average relative |deviation| from true value=
 $1/3*(3.14-2.70 + 3.14-2.71 + 3.14-2.72)/3.14 = 13.7\%$

R_{merge}
formula!

What is the “true value”?

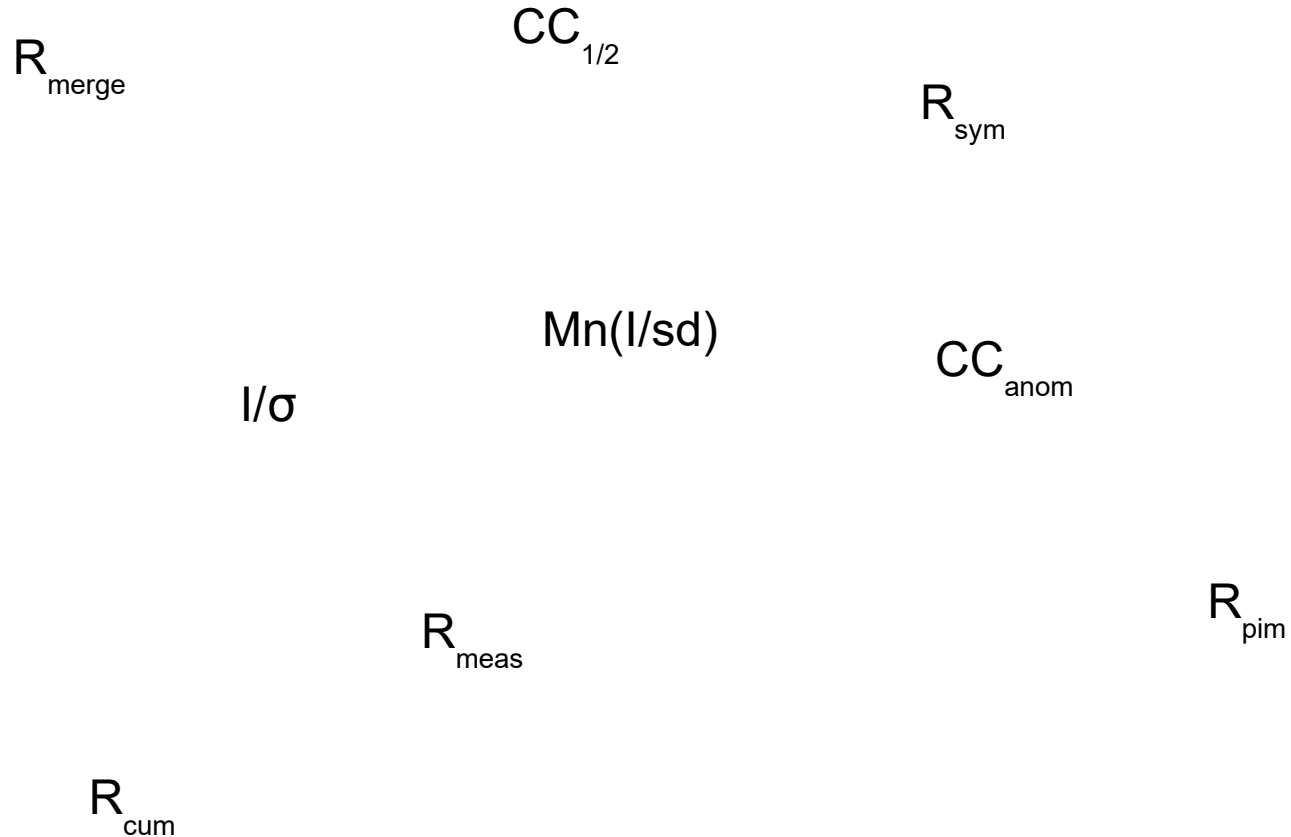
- if only **random error** exists, $\langle \text{accuracy} \rangle = \langle \text{precision} \rangle$
- if unknown **systematic error** exists, true value cannot be found from the data themselves.
- $\langle \text{accuracy} \rangle$ and $\langle \text{precision} \rangle$ differ by the unknown systematic error
- $\langle \text{precision} \rangle$ can easily be calculated, but not $\langle \text{accuracy} \rangle$

All data quality indicators estimate *precision* (only), but YOU (should) want to know *accuracy*!

- **Rules:** “The data processing statistics tells me (and the reviewers!) how good my data are.
To satisfy reviewers, the indicators must be good.”
- **Suboptimal result:** these rules encourage
 - overexposure of crystal to lower R_{merge}
 - data collection “strategy” with low multiplicity
 - statistics massaging: throw away potentially useful data
- **Concepts:**
 - Data processing logfiles report the *precision* (consistency) of the data, *not* their *accuracy* (agreement with truth).
 - averaging increases accuracy *unless* the data repeat systematic errors
 - outliers may be correctly (“true positive”) or incorrectly (“false positive”) identified. Rejections always *increase* precision, but may *decrease* accuracy!

2nd example: confusion by
multitude and properties of
crystallographic indicators

Confusion – what do these mean?



Calculating the precision of unmerged (individual) observations

$\langle I_i / \sigma_i \rangle$ (σ_i from error propagation,
 i =individual measurement)

$$R_{merge} = \frac{\sum_{hkl} \sum_{i=1}^n |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^n I_i(hkl)}$$

$$R_{meas} = \frac{\sum_{hkl} \sqrt{\frac{n}{n-1}} \sum_{i=1}^n |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^n I_i(hkl)}$$

$$R_{meas} \sim 0.8 / \langle I_i / \sigma_i \rangle$$

Calculating the precision of **merged** data

a) using the $\sqrt{n}$ law of error propagation (Wikipedia “weighted arithmetic mean”):

$$\langle I/\sigma(I) \rangle \quad R_{pim} = \frac{\sum_{hkl} \sqrt{\frac{1}{n}} \sum_{i=1}^n |I_i(hkl) - \bar{I}(hkl)|}{\sum_{hkl} \sum_{i=1}^n I_i(hkl)} \quad R_{pim} \sim 0.8 / \langle I/\sigma \rangle$$

b) by comparing averages of randomly selected half-datasets X,Y:

H,K,L	I_i in order of measurement	Assignment to half-dataset	Average I of	
			X	Y
1,2,3	100 110 120 90 80 100	X, X, Y, X, Y, Y	100	100
1,2,4	50 60 45 60	Y X Y X	60	47.5
1,2,5	1000 1050 1100 1200	X Y Y X	1100	1075
...				

Then calculate **Pearson correlation coefficient: $CC_{1/2}$ on X, Y** 12

Measuring the precision of **merged** data with a correlation coefficient

Correlation coefficient $cc_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$ has clear meaning and well-known statistical properties

a) Significance of its value can be assessed by Student's t-test: e.g. $CC > 0.3$ is significant at $p=0.01$ for $n > 100$;
 $CC > 0.08$ is significant at $p=0.01$ for $n > 1000$

b) From $CC_{1/2}$, we can analytically estimate **CC of the merged dataset against the true (unknown) intensities** using $CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$, assuming absence of systematic error.
 $CC^* = \text{upper limit of } CC_{\text{work}} / CC_{\text{free}}$ in refinement (*data quality limits model quality*): $CC_{\text{work}} > CC^*$ implies overfitting
 = model agrees better with data than the true signal does

Rule: “the quality of the data that I use for refinement can be assessed by $R_{\text{merge}}/R_{\text{meas}}$. Data with $R_{\text{merge}}/R_{\text{meas}} > \text{e.g. } 60\%$ are useless.”

- Suboptimal result: Wrong indicator. Wrong high-resolution cutoff. Wrong data-collection strategy. Strong radiation damage.

Concept: - use precision of the *merged* data if you are interested in the suitability of the data for MR, phasing and refinement.

- Like $R_{\text{merge}}/R_{\text{meas}}$, R_{pim} goes to infinity for weak data, whereas $R_{\text{work}}/R_{\text{free}}$ approach a constant: R_{pim} cannot predict model agreement with data
- $\langle I/\sigma \rangle$ or $\langle I \rangle / \langle \sigma \rangle$ - but how to calculate σ ; and which cutoff??
- $CC_{1/2}$, CC^* - no need for σ ; normalized; predicts agreement of data with optimal model

Assmann, G., Wang, M., and Diederichs, K. (2020), Acta Cryst D76, 636:
validation of $CC_{1/2}$ as optimization target

3rd example: *improper* crystallographic reasoning

situation: data to 2.0 Å resolution

using all data: $R_{\text{work}}=19\%$, $R_{\text{free}}=24\%$ (overall)

cut at 2.2 Å resolution: $R_{\text{work}}=17\%$, $R_{\text{free}}=23\%$

- **Rule:** “The lower the R-value, the better.”
„cutting at 2.2 Å is better because it gives lower R-values“
- (Potentially) suboptimal result: throwing away data.
- **Concept:** indicators may only be compared if they refer to the *same* reflections.

Proper crystallographic reasoning

.... requires three concepts:

1. Better data allow to obtain a better model
2. A better model has a lower R_{free} , and a lower $R_{\text{free}} - R_{\text{work}}$ gap
3. *Comparison* of model R-values is only *meaningful* when using the *same* data

Taking these together, this leads us to the „*paired refinement technique*“: compare models in terms of their R-values against the *same* data.

P.A. Karplus and K. Diederichs (2012) Linking Crystallographic Data with Model Quality. *Science* **336**, 1030-1033.

4th ex.: Resolution of the data

Rules:

1. Worst: cutoff based on $R_{\text{merge}}/R_{\text{meas}}$ (which value?)
2. Better: cutoff based on $\langle I/\sigma(I) \rangle$ (which value?) merged data
3. Even better, but not good: cutoff based on $CC_{1/2}$ (which value?)
(some people say 50%, others 30-50%; EM “gold standard” is 14.3%) merged data, no σ

Concepts:

1. “ideally, we would determine the point at which adding the next shell of data is not adding any statistically significant information” (P. Evans)
2. paired refinement method proper comparison
3. only a good model can extract information from weak data external
4. $R_{\text{work}}/R_{\text{free}}$ of model against *noise* is ~43% (G. Murshudov) validation

Advice: be generous at the data processing stage, and
decide only at the very end of refinement
Deposit the data up to the resolution where $CC_{1/2}$ becomes insignificant!

Resolution of the model

Rule:

the resolution of the *model* is the resolution of the data it was refined against

Concepts:

1. the notion “resolution of a model” is misguided – it answers the wrong question!
2. *resolution of a map* (Urzhumtsev *et al*) is well-defined: how far are features apart that we can distinguish? depends on Wilson-B
3. better to ask about precision and accuracy of the model
 - precision: reproducibility of coordinates
 - accuracy: which errors are present? much more important!

Summary

- Crystallographic decisions are often based on *rules* of (if anything) only historical interest. These rules frequently lead to *improper shortcuts* being taken
- “make everything as simple as possible, but not simpler” (attributed to A. Einstein)
- Rules may be needed in expert systems; however, humans should rather learn, apply and further develop the underlying *concepts*

Thank you for your attention!

References:

Karplus, P.A. and Diederichs, K. (2015) Assessing and maximizing data quality in macromolecular crystallography. *Current Opinion in Struct.Biol.* **34**, 60-68.

Diederichs, K. (2015) Crystallographic data and model quality. in: Nucleic Acids Crystallography (Ed. E. Ennifar), *Methods in Molecular Biology* **1320**, 147-173.

Diederichs, K. (2017) Dissecting random and systematic differences between noisy composite data sets. *Acta Cryst.* **D73**, 286-293.

Assmann, G., Wang, M., and Diederichs, K. (2020) Making a difference in multi-data-set crystallography: simple and deterministic data-scaling/selection methods. *Acta Cryst.* **D76**, 636-652

(PDFs at <http://cms.uni-konstanz.de/strucbio/diederichs-group/publications>)