

# Assessing data quality

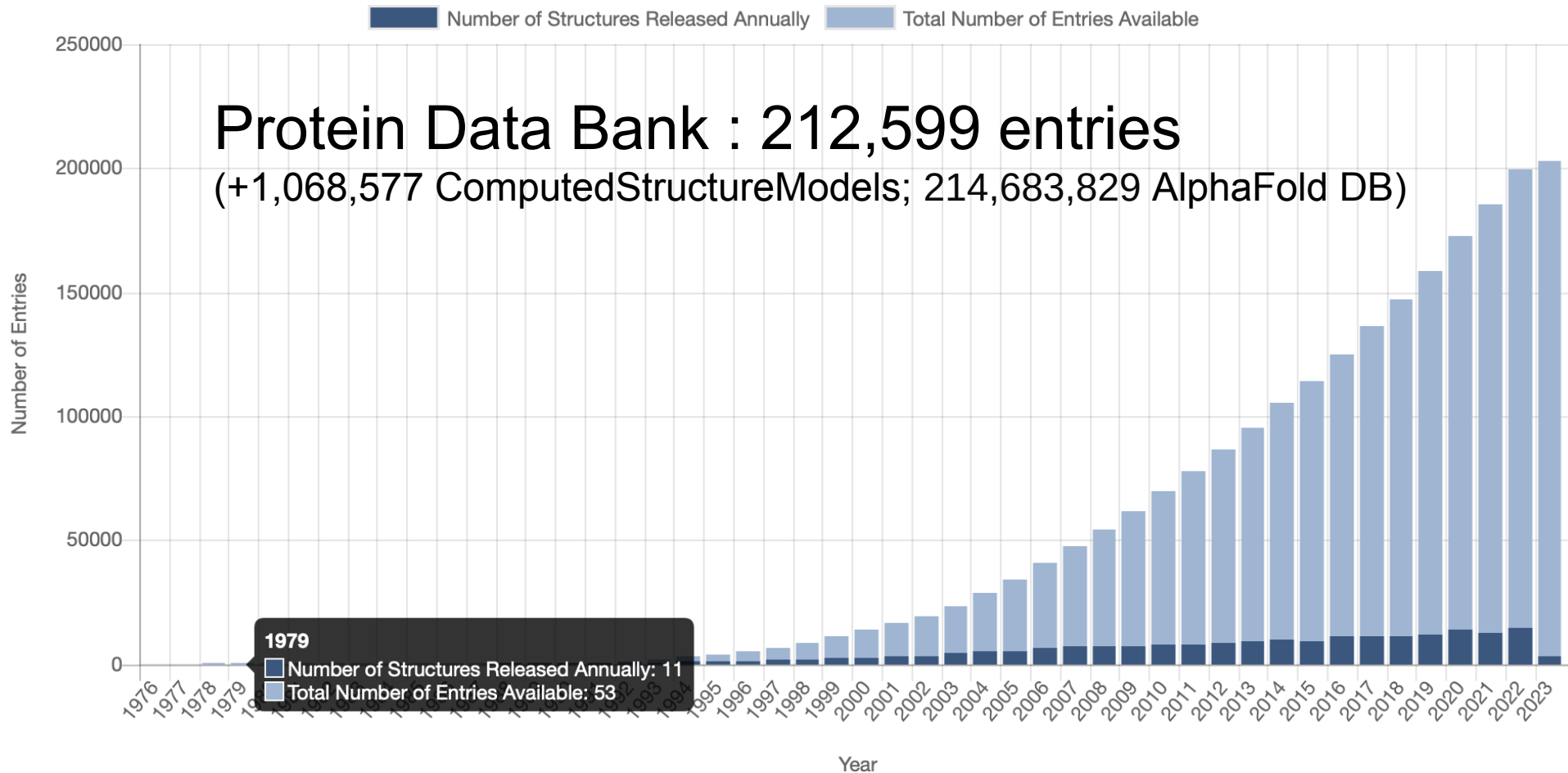
**Kay Diederichs**

`kay.diederichs@uni-konstanz.de`



**Protein Crystallography /  
Molecular Bioinformatics  
University of Konstanz, Germany**

# Crystallography has been extremely successful



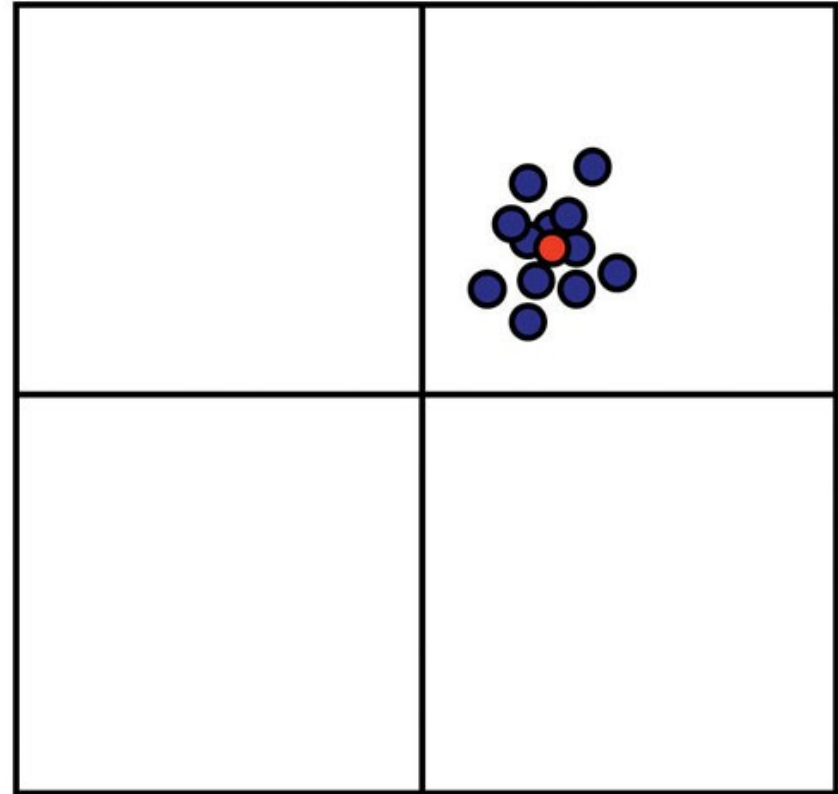
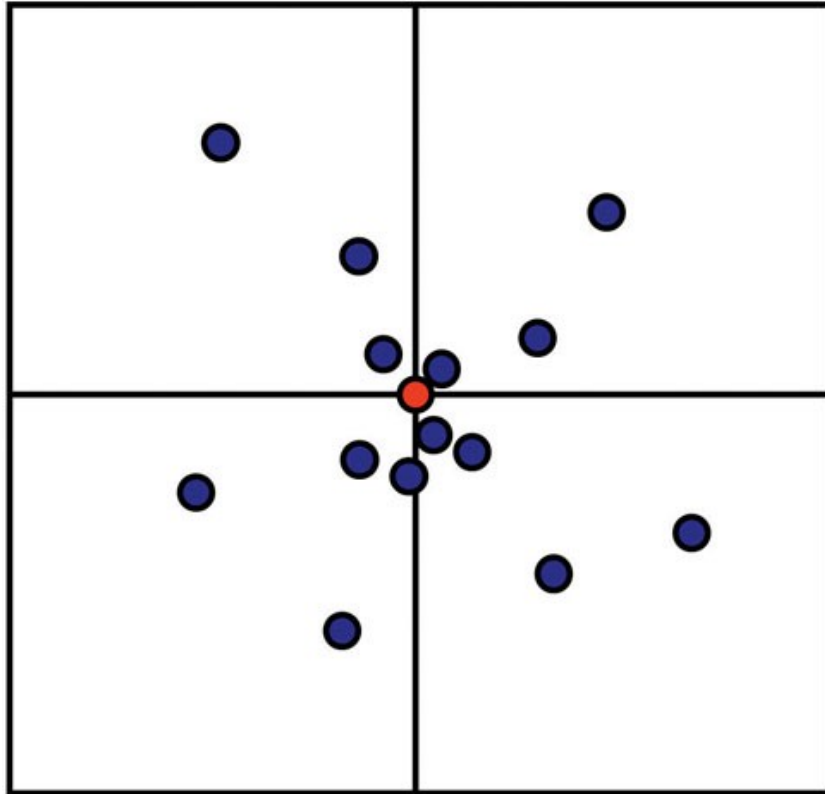
# however ...

- Macromolecular crystallography suffers from *rules* that may have been useful in the past, but are still commonly used today and result in *wrong decisions*, and *misunderstandings*
- another reason for the misunderstandings (and difficulties for those learning and practising crystallography) is that we keep *comparing the wrong units*
- data quality statistics are presented in *confusing* and wrong ways
- *Crystallography is therefore difficult to understand*

1<sup>st</sup> example: let's talk about the difference between, and the relevance of **precision** and **accuracy**



“Quality”



© Garland Science 2010

B. Rupp, Bio-  
molecular  
Crystallography

**Accuracy**  
**Precision**

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

# Numerical example

Repeatedly determine  $\pi=3.14\dots$  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:  
 $=(|3.14-3.1| + |3.14-3.2| + |3.14-3.0|)/(3*3.14) = 2.5\%$

$R_{\text{merge}}$   
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  $\pi=3.14\dots$  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=  
 $(3.14-2.70 + 3.14-2.71 + 3.14-2.72)/(3*3.14) = 13.7\%$

$R_{\text{merge}}$   
formula!

Calculation of precision needs multiple observations of the same quantity; calculation of accuracy needs the true value.

# Consequences

- **Known systematic effects** can usually be made part of the model, or compensated in data processing.
- If **unknown systematic error** exists, the true value cannot be found from the data.
- Without unknown systematic error, **averaging of multiple observations** approximates the true value. Higher multiplicity → better approximation. In that case, only **random error** exists, and  $\langle \text{accuracy} \rangle \sim \langle \text{precision} \rangle$ .
- Accuracy and precision differ by the unknown systematic error. Precision is an **optimistic estimate** of the accuracy!
- Precision can easily be calculated, but not accuracy – because the true value is usually not known.
- *The typical “Table 1” has data quality indicators estimate precision (only), but what we really want to know is accuracy!*
- *Data quality is always worse than what the precision suggests.*

**Rule:** “The data processing statistics tells me (and the reviewers!) how good my data are.  
To satisfy reviewers, the indicators must be good.”

*This is a misunderstanding! Table 1 does not tell the correct story.*

Rules that optimize the precision may compromise the accuracy.

- overexposure of crystal to lower  $R_{\text{merge}}$  (a precision indicator). This increases the radiation damage, and thus the systematic errors.
- throw away potentially useful data, e.g. weak data at high resolution. This may lead to (cosmetically) lower R-values, but worse model.

## **Concepts:**

- averaging increases accuracy *unless* the data *repeat systematic errors*
- rejecting data as outliers always *increases* precision, but may *decrease* accuracy



# Traditional precision indicators focus on the random error

**Random error** is due to

.. the quantum nature of matter and energy: photon counting, and electronic noise in detector, and

.. is proportional to square root of measured value.

***In crystallography, it dominates the error at high resolution.***

**Systematic error** is due to

.. **Crystal**: variation in crystallization conditions, composition, conformation, *radiation damage during experiment*.

.. **Beamline**: shadows, absorption, vibrations, varying photon/electron flux.

.. **Processing software**: inaccurate or incomplete modelling of experiment

.. is proportional to measured value (often 1..10% but sometimes much more e.g. in case of shadows and overloads).

***In crystallography, systematic error dominates at low resolution.***

# There is a single indicator for systematic error in data

- Compare – for a given dataset – the errors of the weak data with those of the strong data. This establishes the so-called “error model”.
- The error model can be analyzed to estimate the  $I/\sigma$  of a (hypothetical) super-strong reflection of this dataset. This is called  $I$ -over- $\sigma$ -asymptotic (**ISa**).
- **ISa depends on the systematic error only** (*not* on crystal size, flux, exposure ...)

Rules of thumb:

<5 something is wrong (e.g. spacegroup or indexing)

5 ..10 marginal data, high systematic error

10 .. 20 more and more useful data

20 .. 30 good data

>30 great data, little systematic error

Programs XDS and newer versions of AIMLESS report ISa.

**Its reciprocal is the percentage of systematic error.**

Diederichs, K. (2010) Quantifying instrument errors in macromolecular X-ray data sets. *Acta Cryst.* **D66**, 733-740.

2<sup>nd</sup> example: confusion by  
multitude and properties of  
crystallographic indicators

# Comparing model and data

During and after refinement, we measure the agreement of model and data, with log-likelihood, and:

- ... with R-values ( $R_{\text{work}}$ ,  $R_{\text{free}}$ )

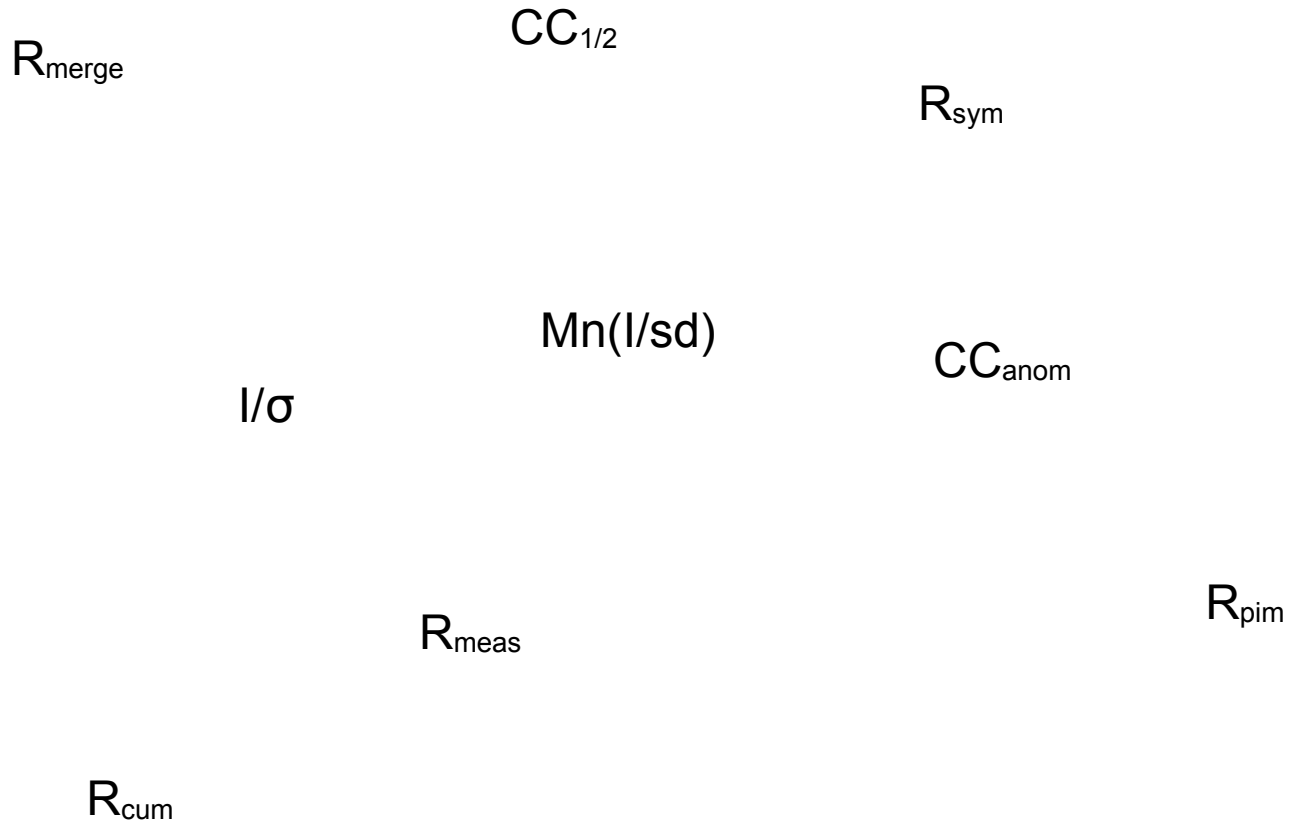
$$R = \frac{\sum_{hkl} |F_{\text{obs}}(hkl) - F_{\text{calc}}(hkl)|}{\sum_{hkl} F_{\text{obs}}(hkl)}$$

- ... or with correlation coefficients ( $CC_{\text{work}}$ ,  $CC_{\text{free}}$ )

$$CC = \frac{\sum_{hkl} (I_{\text{obs}}(hkl) - \overline{I_{\text{obs}}})(I_{\text{calc}}(hkl) - \overline{I_{\text{calc}}})}{\sqrt{\sum_{hkl} (I_{\text{obs}}(hkl) - \overline{I_{\text{obs}}})^2 \sum_{hkl} (I_{\text{calc}}(hkl) - \overline{I_{\text{calc}}})^2}}$$

where the sums go over all unique  $hkl$  values

For data: confusion – what do these  
“Table 1” indicators tell me?



# Calculating the precision of unmerged (individual) observations

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

is biased (Diederichs & Karplus, 1997) → shouldn't be used!

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

$\langle |I_i/\sigma_i| \rangle$  ( $\sigma_i$  from error propagation)

→  $R_{meas} \sim 0.8 / \langle |I_i/\sigma_i| \rangle$  if error estimates internally consistent

averaging/"merging":  $I = \frac{\sum_1^N \frac{I_i}{\sigma_i^2}}{\sum_1^N \frac{1}{\sigma_i^2}}$  and  $\sigma = \sqrt{\frac{1}{\sum_1^N \frac{1}{\sigma_i^2}}}$

# Calculating the precision of **merged** data

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

$$\langle I/\sigma(I) \rangle$$

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

$$R_{pim} \sim 0.8 / \langle I/\sigma \rangle$$

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

| H,K,L | $I_i$ of individual measurements | 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

# 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: could this CC arise by chance (random data)? Typically, call **significant if low likelihood** (1% or 0.1%).

b) From  $CC_{1/2}$ , we can analytically estimate **CC\* = correlation of merged dataset against the true intensities** using assuming absence of systematic error.

$$CC^* = \sqrt{\frac{2CC_{1/2}}{1 + CC_{1/2}}}$$

c)  $CC_{\text{work/free}}$  and  $CC^*$  are intensity-based → meaningful comparison!

d) If  $CC_{\text{work}} > CC^*$  then this implies overfitting (because model agrees better with data than the true signal does). This means that  $CC_{\text{work}}/CC_{\text{free}}$  in refinement is limited by  $CC^*$  : **data quality limits model quality**



# Comparing model R-values with data R-values

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

**Misunderstanding:** This is the wrong indicator!

- A model is refined against *merged data*;  $R_{\text{merge}}$  is for *unmerged data*!
- Model R-values  $R_{\text{work/free}}$  are based on *amplitudes*, data R-values  $R_{\text{merge}}, R_{\text{pim}}$  on *intensities*
- $R_{\text{merge}}, R_{\text{pim}}$  go to infinity for weak data, whereas  $R_{\text{work}}/R_{\text{free}}$  approach a constant ( $\sim 60\%$ ). R-values do not predict model agreement with data  
→ **model and data R-values are not comparable!**

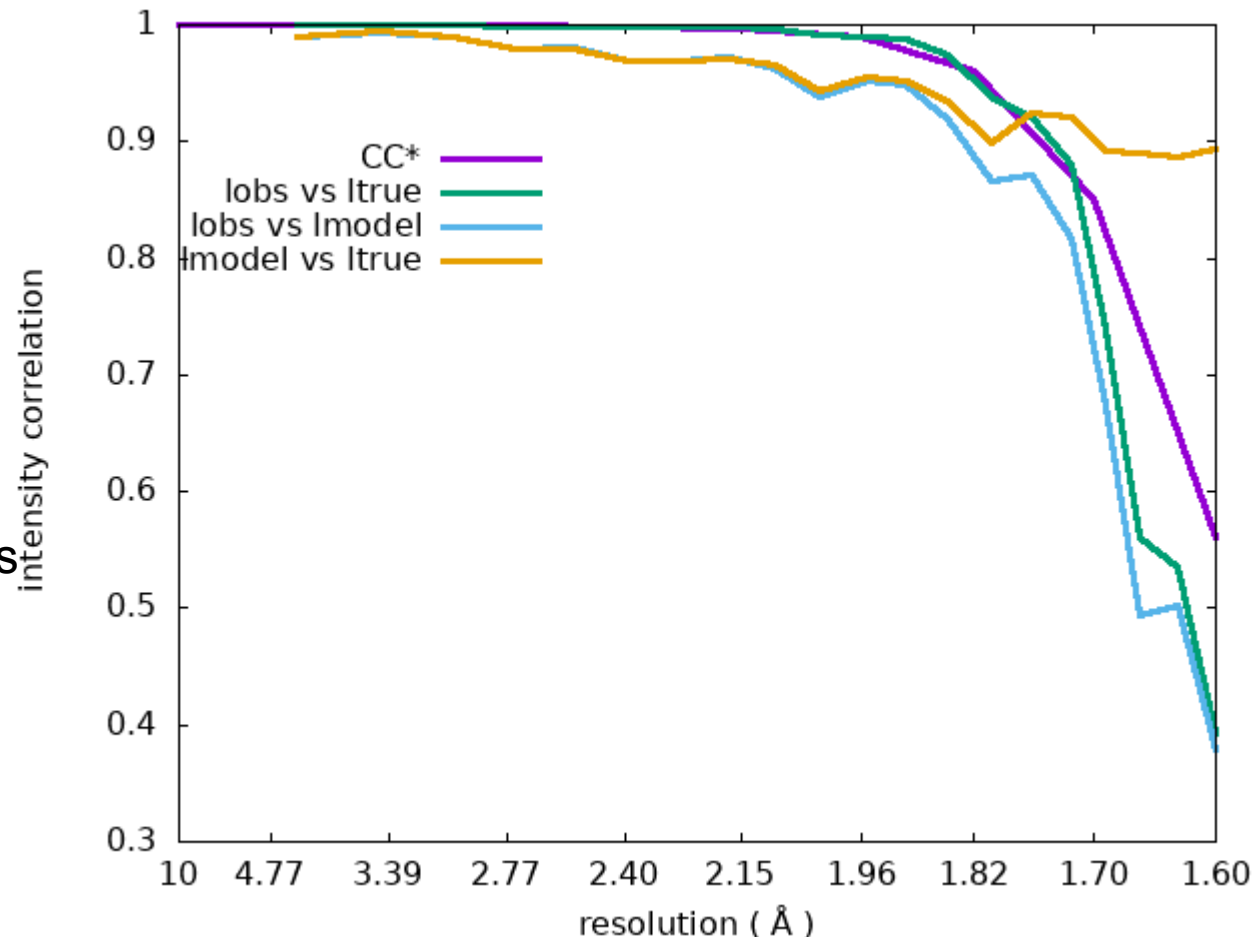
**Resulting mistakes:** Wrong high-resolution cutoff, wrong data-collection strategy, strong radiation damage, ...

# How to estimate accuracy of data?

- **optimistic** estimate:  
precision of data,  $CC^*$
- **exact** – only known for  
simulated data,  $CC(I_{\text{obs}}, I_{\text{true}})$
- **pessimistic** estimate:  
compare model with data,  $CC(I_{\text{obs}}, I_{\text{model}})$

## Insights:

- neglect of systematic error is what makes  $CC^*$  optimistic
- model deficiencies are responsible for making  $CC(I_{\text{obs}}, I_{\text{model}}) < CC(I_{\text{obs}}, I_{\text{true}})$
- $CC(I_{\text{model}}, I_{\text{true}})$  shows that  $I_{\text{model}}$  is more accurate than  $I_{\text{obs}}$  at high resolution!
- model accuracy is underestimated if evaluated from  $CC(I_{\text{obs}}, I_{\text{model}})$  !



$I_{\text{true}}$  is available since these are simulated data (Diederichs, 2009)

$CC$  = “correlation coefficient on  $I$ ” is useful as a “common currency” for comparing agreement values!

# 3<sup>rd</sup> 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:  $R_{\text{work}}=17\%$ ,  $R_{\text{free}}=23\%$

- **(Wrong) comparison:** “The lower the R-value, the better.”  
→ „cutting at 2.2 Å is better: it gives lower R-values“  
→ (Potentially) bad result: throwing away data.
- **Correct question:** which model is better? (the goal of refinement is to optimize the model, not the R-values!)
- **Insight:** indicators may only be compared if they refer to the *same* reflections.

# Proper comparison

“ideally, we would determine the point at which adding the next shell of data is not adding any statistically significant **information**” (Evans, 2011)

## How to do it:

- use unchanged higher-res model for  $F_{\text{calc}}$  but compare R-values of only the lower-res reflections against those of the lower-res model
- the better model has the lower  $R_{\text{free}}$ , and the lower  $R_{\text{free}}-R_{\text{work}}$  gap
- can be repeated for different high-res cutoffs
- „*paired refinement technique*“ (Karplus & Diederichs, 2012).
- available in PDB-REDO and PAIREF (standalone or CCP4)

# Resolution of the data

## Questionable comparisons:

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  $\sigma$

## Solution:

1. 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% (Murshudov, 2011) validation

## Advice:

- be generous at the data processing stage
- decide only at the very end of refinement about the resolution cutoff
- deposit the data up to the resolution where  $CC_{1/2}$  becomes insignificant!

# Summary

- Crystallographic statistics are plagued by unsuitable indicators (e.g.  $R_{\text{merge}}$ ) whose properties are misunderstood, but which are perpetuated and enshrined in “*Table 1*”.
- One confusion is the *mix-up of precision and accuracy*.
- Comparing model R-values ( $R_{\text{work}}, R_{\text{free}}$ ) to data R-values ( $R_{\text{merge}}, R_{\text{pim}}$ ) is not sensible. Historically, this attempt has lead to confusing and wrong rules and decisions. Conversely, correlation coefficients as a function of resolution can be meaningfully compared, and interpreted.
- Yet another source of confusion is the attempt to compare model R-values ( $R_{\text{work}}, R_{\text{free}}$ ) referring to *different* sets of reflections. This leads to recurring discussions about resolution cutoffs. The way forward is *paired refinement*.

# Thank you for your attention!

## References:

- Diederichs, K. (2009) Simulation of X-ray frames from macromolecular crystals using a ray-tracing approach. *Acta Cryst.* **D65**, 535-542
- Diederichs, K. (2010) Quantifying instrument errors in macromolecular X-ray data sets. *Acta Cryst.* **D66**, 733-740.
- Evans, P. (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Cryst.* **D67**, 282-292.
- Murshudov, G.N. (2011) Some properties of crystallographic reliability index - Rfactor: effect of twinning. *Appl. Comput. Math.* **10**, 250-261.
- Karplus, P.A. and Diederichs, K. (2012) Linking crystallographic model and data quality. *Science* **336**, 1030-1033.
- 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.
- Malý, M., Diederichs, K., Dohnálek, J., Kolenko, P. (2020) Paired refinement under the control of PAIREF. *IUCrJ* **7**, 681-692.