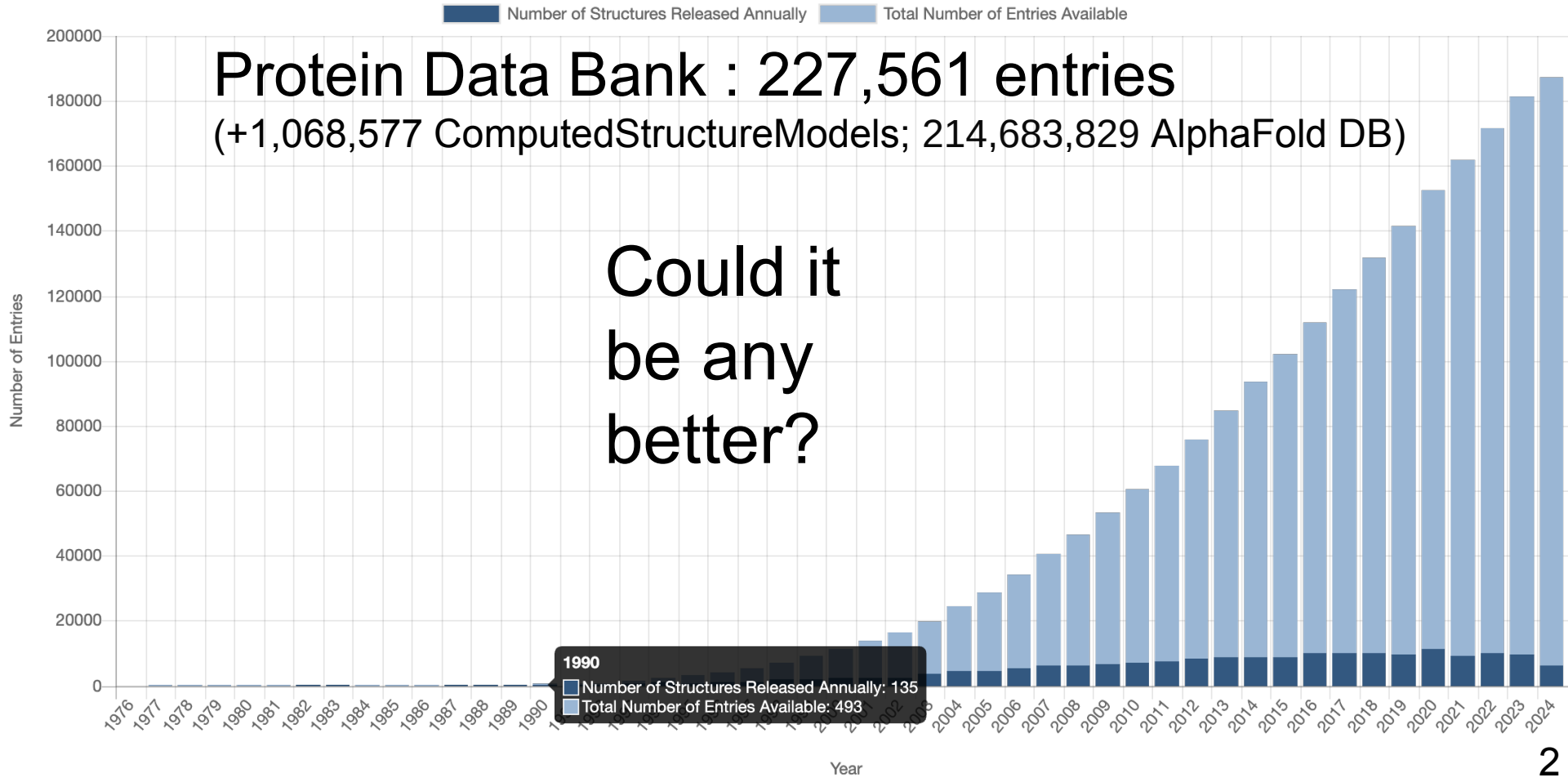
# Assessing crystallographic data quality

## Let's talk about errors, and mistakes

Kay Diederichs
kay.diederichs@uni-konstanz.de


CCP4@DLS 2024-11-26

# Crystallography has been extremely successful

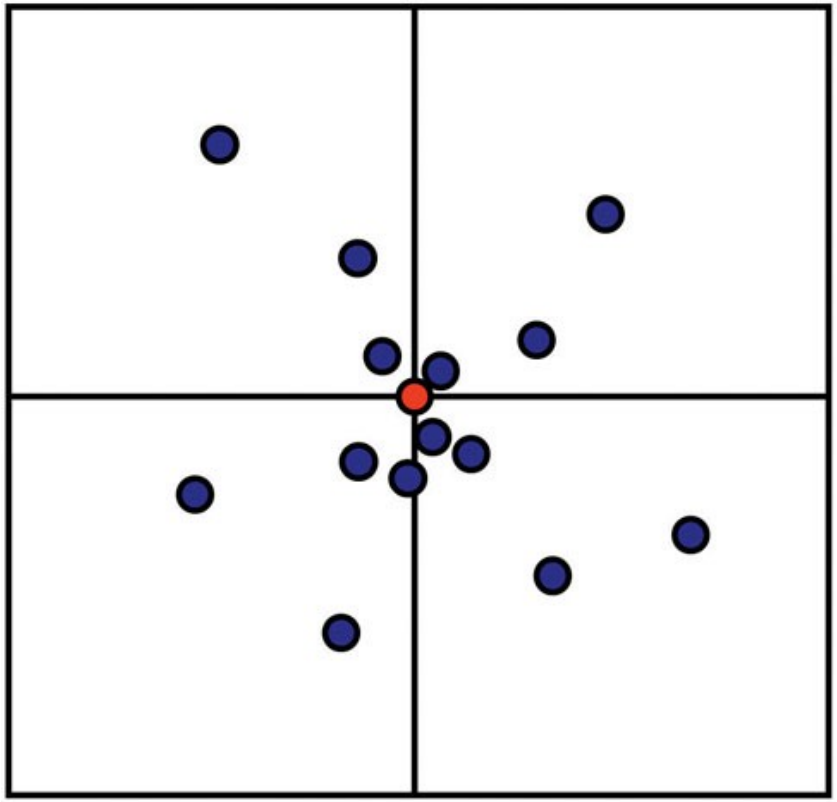


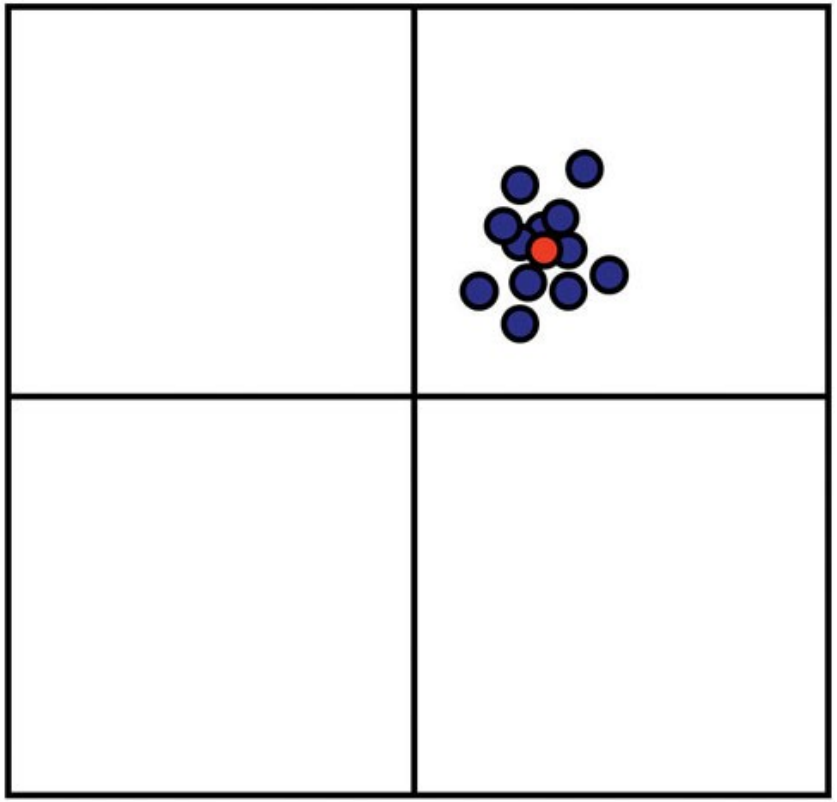
# However ...

- *Crystallography is difficult to understand. Why?*
- 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 indicators*
- Data quality statistics are presented in *confusing* and wrong ways

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


# "Quality"





© Garland Science 2010

**Accuracy** – how different from the *true value*?

**Precision** – how different are *measurements from each other*?

B. Rupp, Biomolecular Crystallography

# Numerical example

Repeatedly determine π=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= relative |deviation from true value|:
=(|3.14-3.1| + |3.14-3.2| + |3.14-3.0|)/(3*3.14) = 2.5%

$R_{merge}$ formula!

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

*pink = data-internal*

Repeatedly determine π=3.14... as 2.70, 2.71, 2.72 :
observations have high precision, low accuracy.

Precision=
(0.01+0+0.01)/(2.70+2.71+2.72) = 0.24%

Accuracy=
(3.14-2.70 + 3.14-2.71 + 3.14-2.72)/(3*3.14) = 13.7%

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

# Deviations from true/targeted value

- **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. **Averaging of multiple observations** may or may not help.
- If only **random error** exists, **averaging of multiple observations** leads to approximation of the true value. Higher multiplicity → better approximation. In that case, <accuracy> ~ <precision> . This is **not true for systematic error**.
- 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" shows data quality indicators estimating precision (only), but what we really want to know is accuracy!*
- *Data accuracy is always worse than what the precision suggests, due to systematic errors that we don't know about.*

# Traditional precision indicators focus on the random error

**Random error** is due to

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

**.. is proportional to square root of measured value.**

***In crystallography, random error 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/σ 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 (e.g. spacegroup or indexing) is wrong (unless Electron Diffraction)
  5 ..10 marginal data, high systematic error
  10 .. 20 more and more useful data
  20 .. 30 good data
  >30 great data, little systematic error

XDS, XSCALE, AIMLESS and DIALS.SCALE report ISa.
**Its reciprocal is the fraction of systematic error in the data.**
Example: ISa = 20 corresponds to 5% of systematic error.

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

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

# Comparing model and data

During and after refinement, we measure the agreement of model and data:

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

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

*blue=model-vs-data*

- … or with correlation coefficients ($CC_{work}$, $CC_{free}$)

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

where the sums go over all unique *hkl* values

# For data: confusion – what do these "Table 1" indicators tell me?

$R_{merge}$

$CC_{1/2}$

$R_{sym}$

Mn(I/sd)

$CC_{anom}$

I/σ

$R_{pim}$

$R_{meas}$

$R_{cum}$

# 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)

→ $R_{meas} \approx 0.8 / \langle I_i/\sigma_i \rangle$ 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 $\sigma = \sqrt{\frac{1}{\sum_1^N \frac{1}{\sigma_i^2}}}$ (Wikipedia "weighted arithmetic mean")

# Calculating the precision of merged data

a) using the √n law of error propagation :

<I/σ(I)>

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

$R_{pim}$ ~ 0.8 / <I/σ>

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**

It can be shown that $CC_{1/2} \approx \frac{1}{1 + \frac{2}{\langle I/\sigma \rangle^2}}$

e.g. <I/σ>=1 → $CC_{1/2}$ ≈ 33%

<I/σ>=2 → $CC_{1/2}$ ≈ 67%

# 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

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

**CC* = correlation(merged dataset , true intensities)**

assuming absence of systematic error.

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

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

# Compare model R-values w/ data R-values??

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

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

- A model is refined against *merged data*; $R_{merge}$ is for *unmerged data*!
- Model R-values $R_{work/free}$ are based on *amplitudes*, data R-values $R_{merge}$,$R_{pim}$ on *intensities*
- $R_{merge}$,$R_{pim}$ go to infinity for weak data, whereas $R_{work}$/$R_{free}$ approach a constant (~60%). Data R-values thus 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, ...

# 3rd example: *improper* crystallographic reasoning

# An example

situation: data to 2.0 Å resolution

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

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

# „Paired refinement technique“

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

- calculate $F_{calc}$ from unchanged **higher-res model** but **compare** R-values of only the lower-res reflections against those of the **lower-res model**
- the better model has the lower $R_{free}$, and the lower $R_{free}$-$R_{work}$ gap
- can be repeated for different high-res cutoffs
- available in PDB-REDO and PAIREF (standalone or CCP4)
- leads to a logical **decision about the high-resolution cutoff**

Karplus, P.A. and Diederichs, K. (2012) Linking crystallographic model and data quality. *Science* **336**, 1030-1033.

Malý, M., Diederichs, K., Dohnálek, J., Kolenko, P. (2020) Paired refinement under the control of PAIREF. *IUCrJ* **7**, 681-692. **https://pairef.fjfi.cvut.cz**

Highly controversial?

# 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: DPI; reproducibility of coordinates
   - accuracy: which errors are present?

# Summary

- Crystallographic statistics are plagued by indicators (e.g. $R_{merge}$) 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 ($R_{work}$,$R_{free}$) to data R-values ($R_{merge}$, $R_{pim}$) 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 ($R_{work}$,$R_{free}$) 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.