Data quality indicators

Kay Diederichs

Crystallography has been highly successful

RCSB Protein Data Bank

rcsb.org/pdb/home/home.do

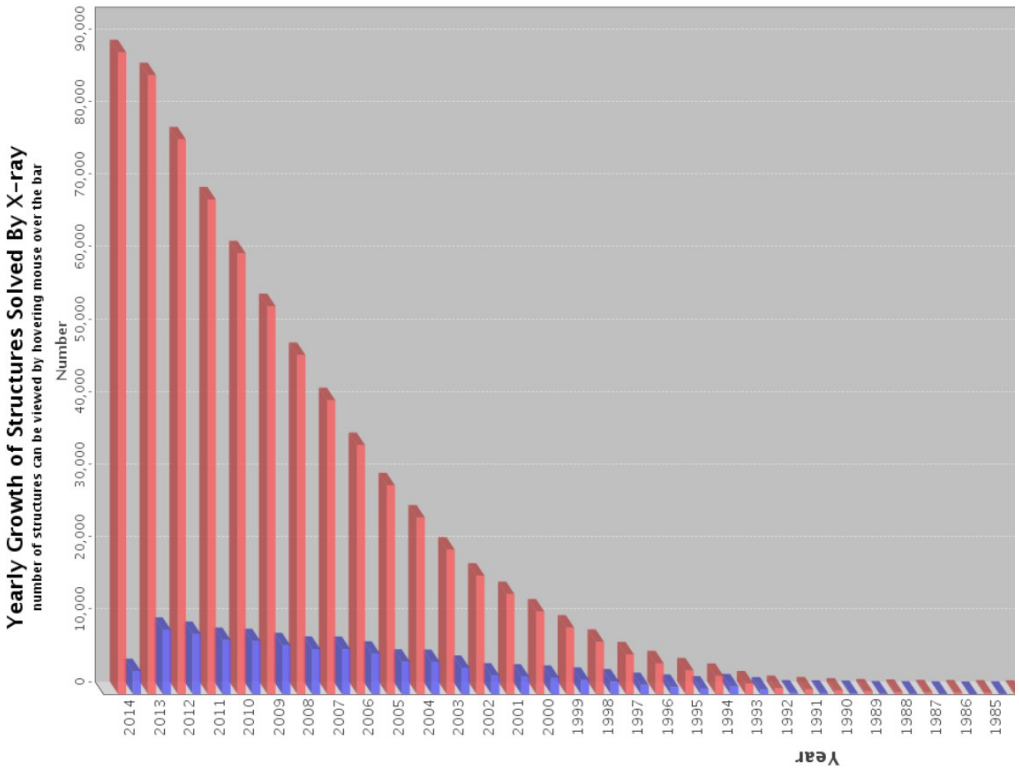
RCSB PDB PROTEIN DATA BANK — PDB-101

A MEMBER OF THE PDB | EMDataBank

An Information Portal to Biological Macromolecular Structures

As of Tuesday May 06, 2014 at 5 PM PDT there are 99928 Structures | PDB Statistics

Now 105839



Could it be any better?

Confusion – what do these mean?

R_{merge}

$CC_{1/2}$

R_{sym}

Mn(I/sd)

CC_{anom}

I/σ

R_{meas}

R_{pim}

R_{cum}

Topics

Signal *versus* noise

Random *versus* systematic error

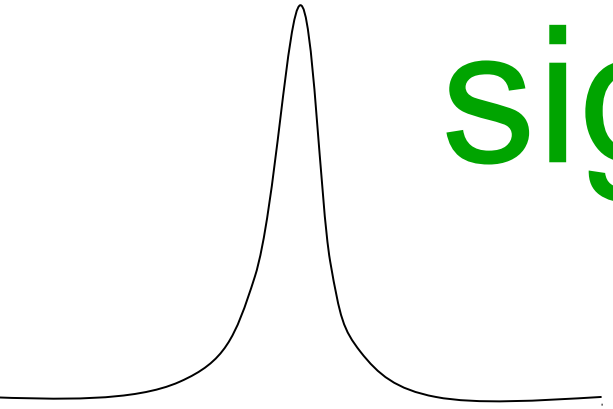
Accuracy *versus* precision

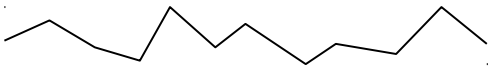
Unmerged *versus* merged data

R-values *versus* correlation coefficients

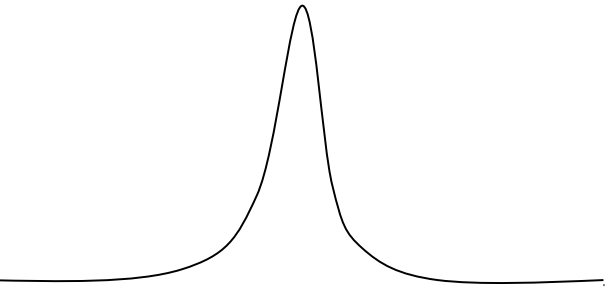
Choice of high-resolution cutoff

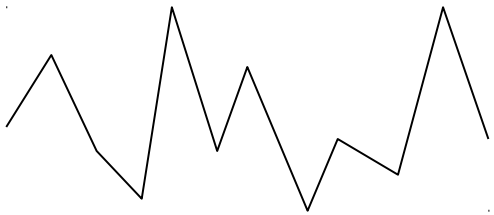
signal *vs* noise





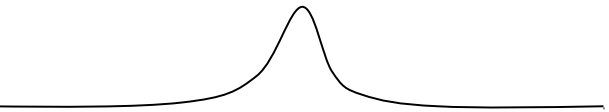
easy

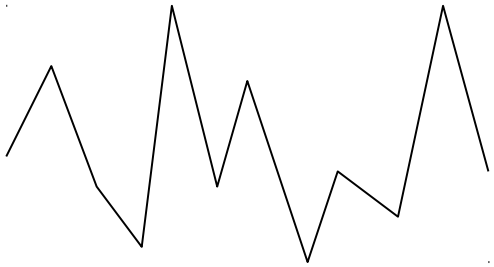




hard

threshold of "solvability"





impossible

James Holton slide

„noise“: what is noise? what kinds of errors exist?

noise = random error + systematic error

random error results from quantum effects

systematic error results from everything else: technical or other macroscopic aspects of the experiment

Random error (noise)

Statistical events:

- photon emission from xtal
- photon absorption in detector
- electron hopping in semiconductors (amplifier etc)

non-obvious

Systematic errors (noise)

- beam flicker (instability) in flux or direction
- shutter jitter
- vibration due to cryo stream
- split reflections, secondary lattice(s)
- absorption from crystal and loop
- radiation damage
- detector calibration and inhomogeneity; overload
- shadows on detector
- deadtime in shutterless mode
- imperfect assumptions about the experiment and its geometric parameters in the processing software
- ...

non-obvious

Adding noise

$$1^2 + 1^2 = 1.4^2$$

$$3^2 + 1^2 = 3.2^2$$

$$10^2 + 1^2 = 10.05^2$$

$$\sigma_1^2 + \sigma_2^2 = \sigma_{total}^2$$

James Holton slide

This law is only valid if errors are independent!

non-obvious

How do random and systematic *error* depend on the *signal*?

random error obeys *Poisson statistics*
error = square root of signal

Systematic error is *proportional* to signal
error = x * signal (e.g. x=0.02 ... 0.10)

(which is why James Holton calls it „fractional error“; there are exceptions)

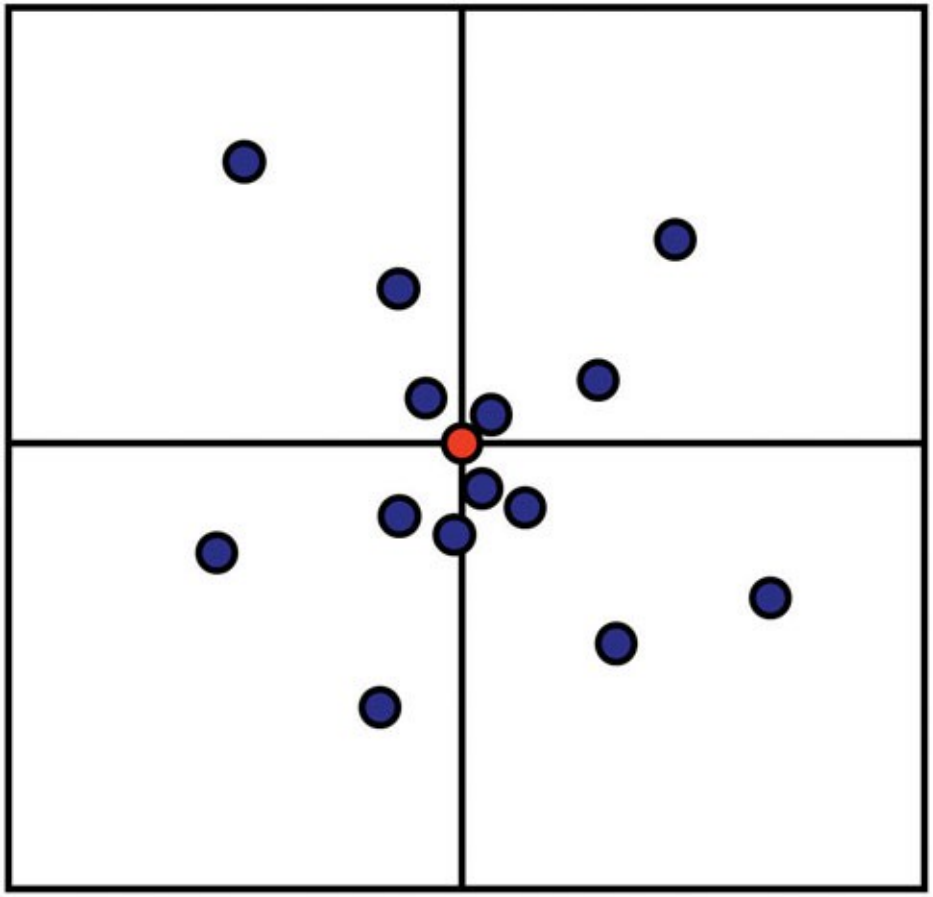
non-obvious

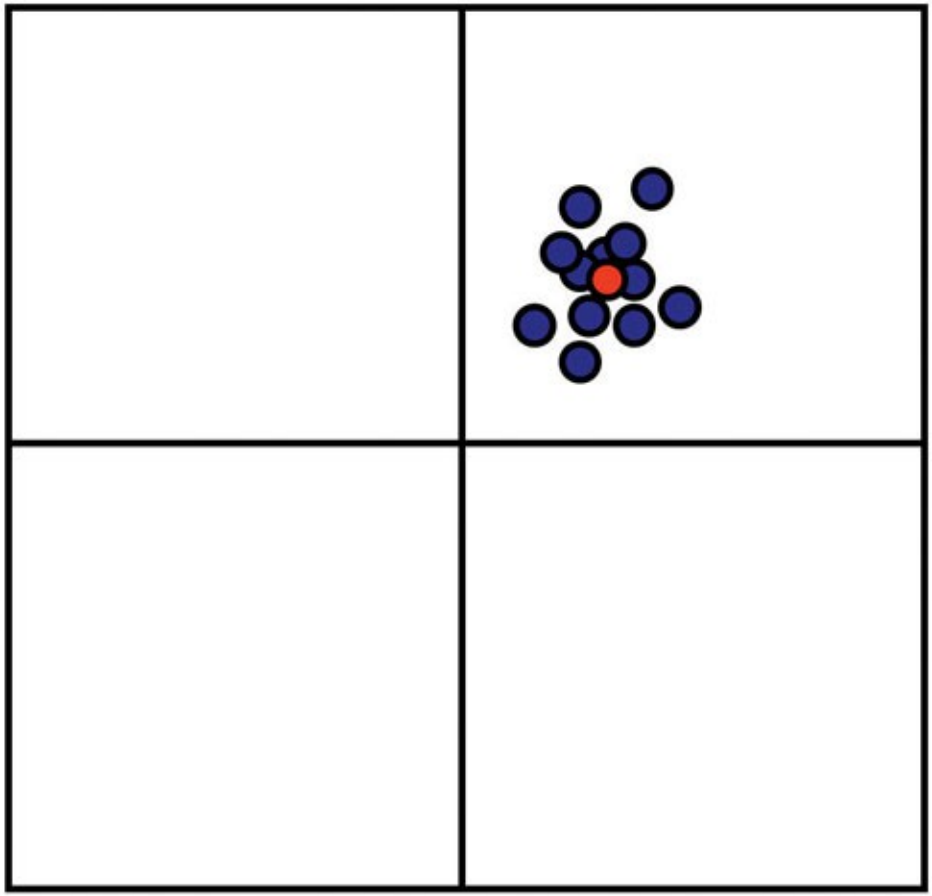
Consequences

- need to add both types of errors
- at high resolution, random error dominates
- at low resolution, systematic error dominates
- but: radiation damage influences both the low and the high resolution (the factor x is low at low resolution, and high at high resolution)

non-obvious

How to measure quality?





B. Rupp, Biomolecular Crystallography

Accuracy – how close to the true value?

Precision – how close are measurements?

non-obvious

What is the „true value“?

- **if only random error exists, accuracy = precision (on average)**
- if unknown systematic error exists, true value cannot be found from the data themselves
- a good model can provide an approximation to the truth
- model calculations do provide the truth
- consequence: precision can easily be calculated, but not accuracy
- **accuracy and precision differ by the unknown systematic error**

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

Numerical example

Repeatedly determine π=3.14159... as 2.718, 2.716, 2.720 :

high precision, low accuracy.

Precision= relative deviation from average value=
(0.002+0+0.002)/(2.718+2.716+2.720) = 0.049%

Accuracy= relative deviation from true value=
(3.14159-2.718) / 3.14159 = 13.5%

Repeatedly determine π=3.14159... as 3.1, 3.2, 3.0 :

low precision, high accuracy

Precision= relative deviation from average value=
(0.04159+0+0.05841+0.14159)/(3.1+3.2+3.0) = 2.6%

Accuracy= relative deviation from true value: 3.14159-3.1 = 1.3%

Calculating the precision of unmerged data

Precision indicators for the unmerged (individual) observations:

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

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

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

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

Averaging („merging“) of observations

Intensities:

$$I = \sum I_i/\sigma_i^2 \; / \sum 1/\sigma_i^2$$

Sigmas:

$$\sigma^2 = 1 \; / \sum 1/\sigma_i^2$$

(see Wikipedia: „weighted mean“)

non-obvious

Merging of observations may improve accuracy and precision

- Averaging („merging“) requires multiplicity („redundancy“)
- (Only) **if errors are unrelated, averaging with multiplicity n decreases the error of the averaged data by sqrt(n)**
- Random errors *are* unrelated by definition: averaging always decreases the random error of merged data
- Averaging *may* decrease the systematic error in the merged data. This requires sampling of its possible values - „true multiplicity“
- **If errors are related, precision improves, but not accuracy**

Calculating the precision of merged data

- using the sqrt(n) law: <I/σ(I)>

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

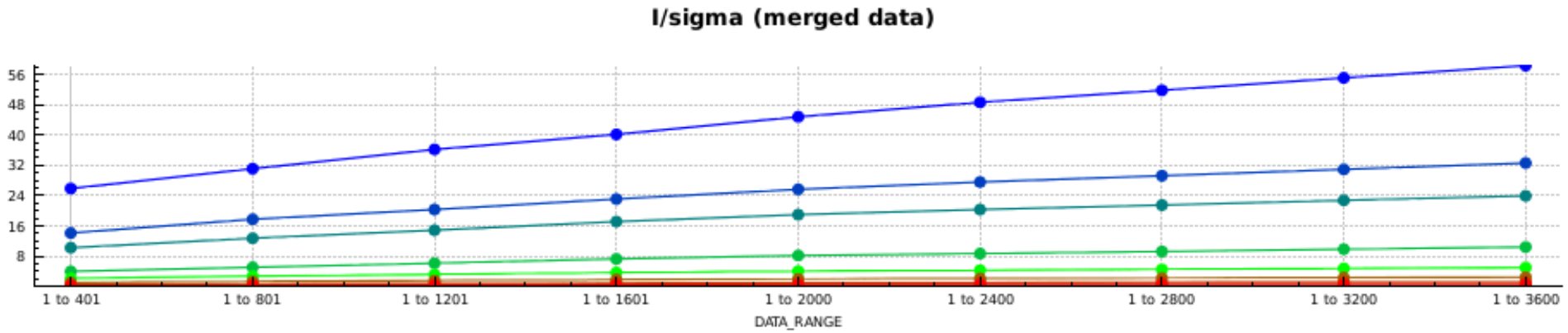
R_{pim} ~ 0.8 / <I/σ >

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

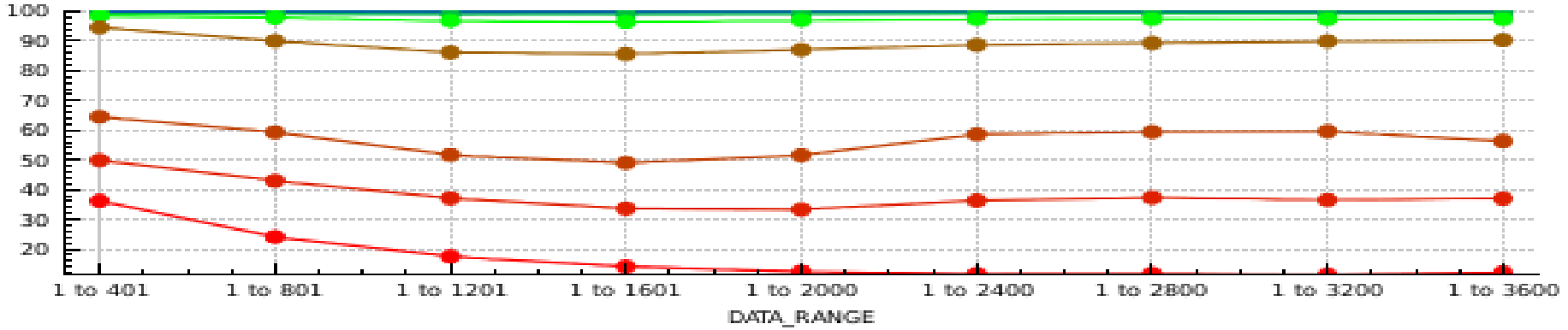
H,K,L	I_i in order of measurement	Assignment to half-dataset	Average I of X	Y
1,2,3	100 110 120 90 80 100	X, X, Y, X, Y, Y	100	100
1,2,4	50 60 45 60	Y X Y X	60	47.5
1,2,5	1000 1050 1100 1200	X Y Y X	1100	1075
...				

(calculate the R-factor (D&K1997) or correlation coefficient (K&D 2012) on X, Y)

I/σ with $\sigma^2 = 1 / \sum 1/\sigma_i^2$



$$r = \frac{\sum_{i=1}^{n}\left(\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)\right)}{\sqrt{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 \sum_{i=1}^{n}\left(y_i - \bar{y}\right)^2}}$$



Shall I use an indicator for precision of *unmerged* data, or of *merged* data?

It is essential to understand the difference between the two types, but you don't find this in the papers / textbooks!

Indicators for precision of *unmerged* data help to e.g.
* decide between spacegroups
* calculate amount of radiation damage (see XDS tutorial)

Indicators for precision of *merged* data assess suitability
* for downstream calculations (MR, phasing, refinement)

Crystallographic statistics - which indicators are being used?

- Data R-values: $R_{pim} < R_{merge} = R_{sym} < R_{meas}$

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

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

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

- Model R-values: R_{work}/R_{free}

$$R_{work/free} = \frac{\sum_{hkl} |F_{obs}(hkl) - F_{calc}(hkl)|}{\sum_{hkl} F_{obs}(hkl)}$$ merged data

- I/σ (for *unmerged* or *merged* data !)
- $CC_{1/2}$ and CC_{anom} for the *merged* data

Decisions and compromises

Which high-resolution cutoff for refinement?

Higher resolution means better accuracy and maps

But: high resolution yields high R_{work}/R_{free}!

Which datasets/frames to include into scaling?

Reject negative observations or unique reflections?

The reason why it is difficult to answer “R-value questions” is that no proper mathematical theory exists that uses absolute differences; concerning the use of R-values, Crystallography is disconnected from mainstream Statistics

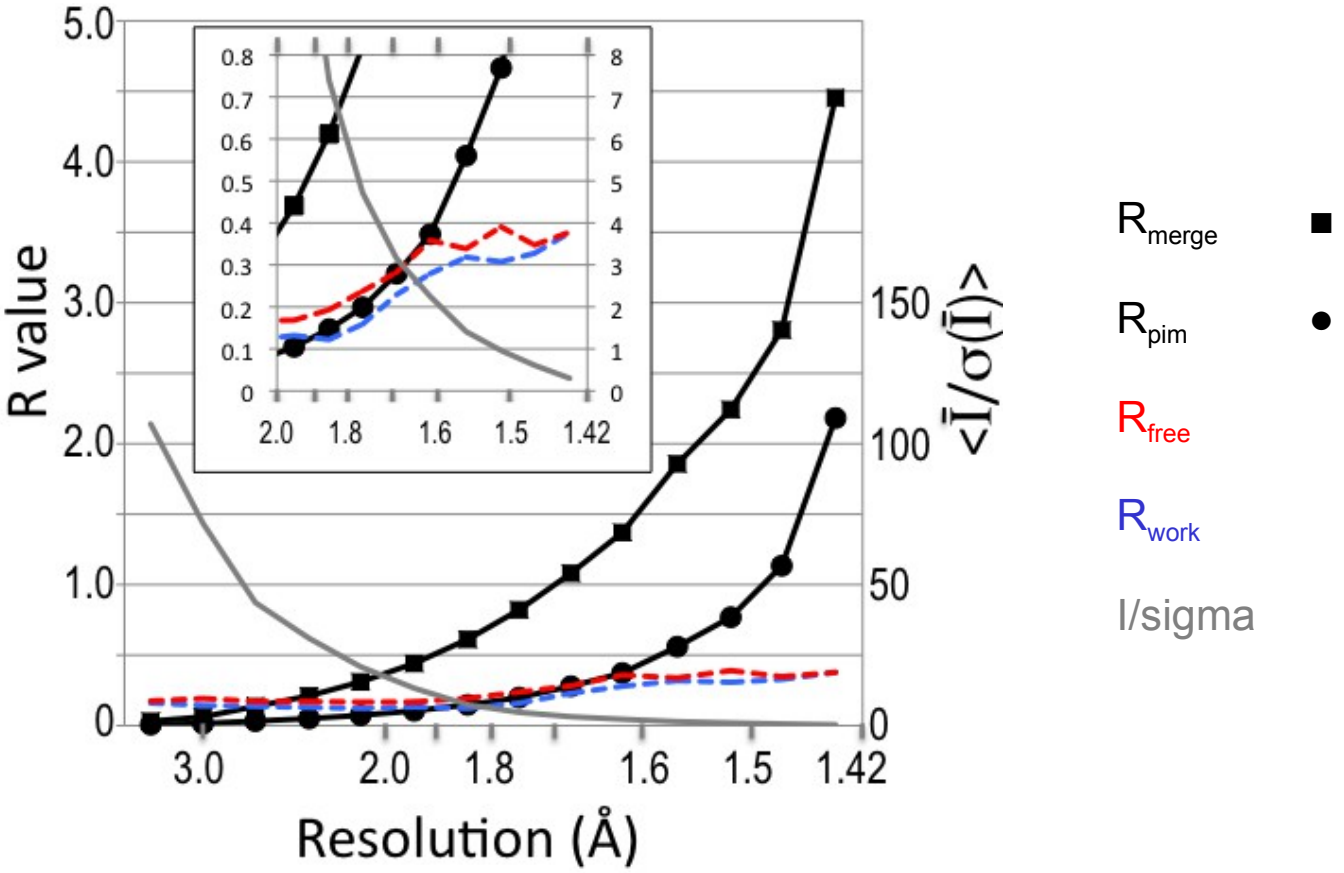
Improper crystallographic reasoning

- typical example: data to 2.0 Å resolution
- using all data: R_{work}=19%, R_{free}=24% (overall)
- cut at 2.2 Å resolution: R_{work}=17%, R_{free}=23%
- „cutting at 2.2 Å is better because it gives lower R-values“

Proper crystallographic reasoning

1. Better data allow to obtain a better model
2. A better model has a lower R_{free}, and a lower R_{free}-R_{work} gap
3. *Comparison* of model R-values is only *meaningful* when using the *same* data
4. Taken together, this leads to the *„paired refinement technique“*: compare models in terms of their R-values against the *same* data.

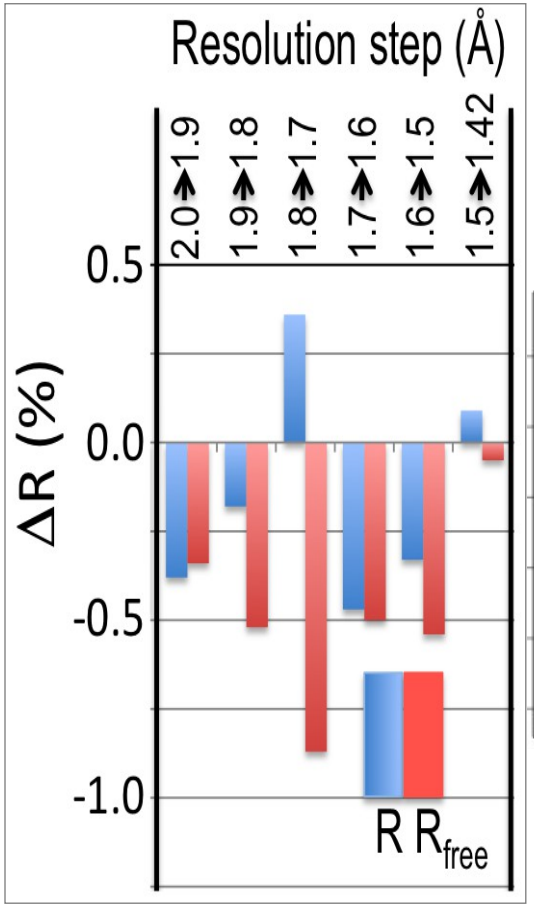
Example: Cysteine DiOxygenase (CDO; PDB 3ELN) re-refined against 15-fold weaker data

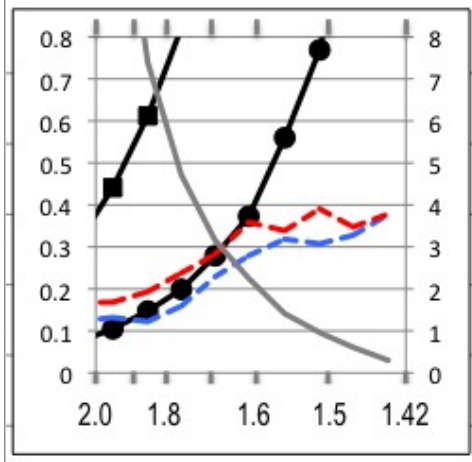


Is there information beyond the conservative hi-res cutoff?

"**Paired refinement technique**":

- refine at (e.g.) 2.0Å and at 1.9Å using the *same* starting model and refinement parameters
- since it is *meaningless* to compare R-values at *different* resolutions, calculate the overall *R-values of the 1.9Å model at 2.0Å* (`main.number_of_macro_cycles=1 strategy=None fix_rotamers=False ordered_solvent=False`)
- $\Delta R = R_{1.9}(2.0) - R_{2.0}(2.0)$





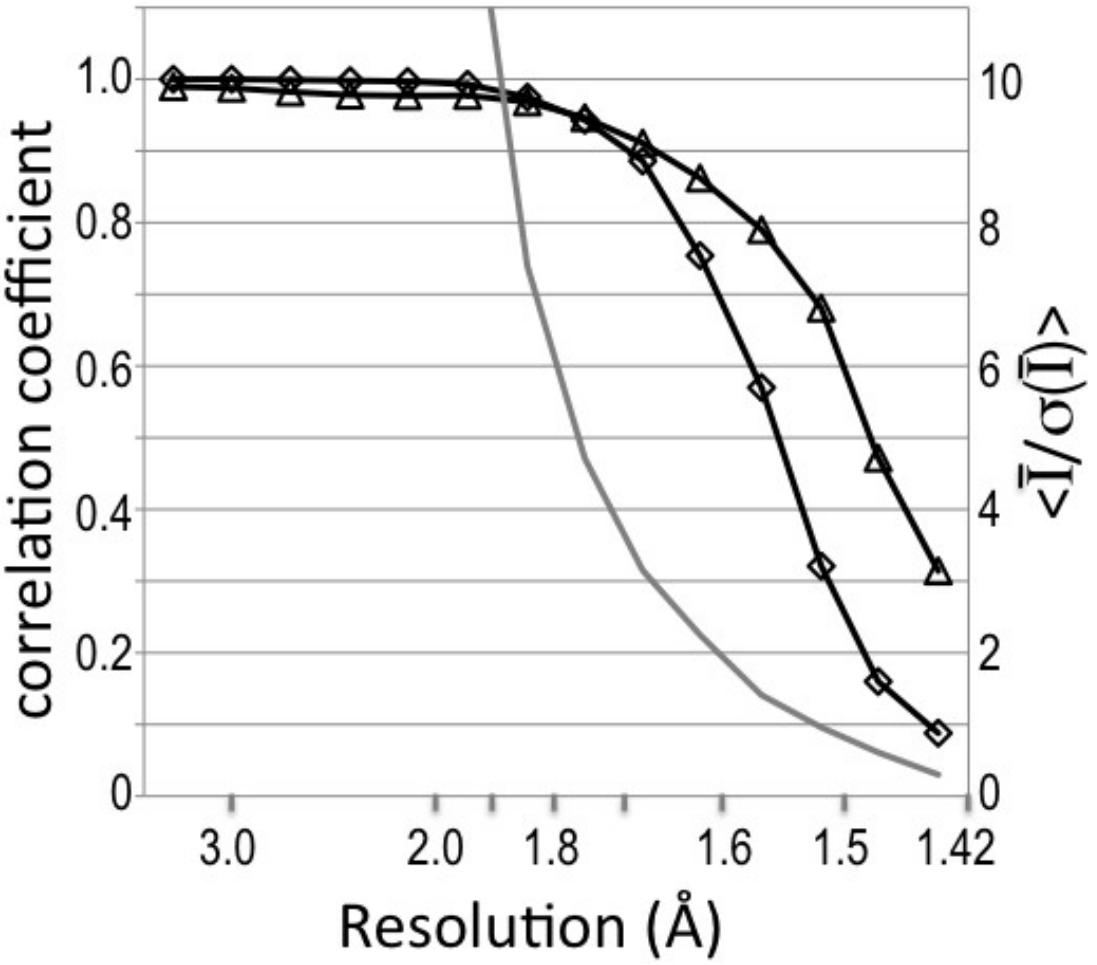
Measuring the precision of merged data with a correlation coefficient

- Correlation coefficient has clear meaning and well-known statistical properties
- 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)
- Apply this idea to crystallographic intensity data: use "random half-datasets" → $CC_{1/2}$ (called CC_Imean by SCALA/aimless, now $CC_{1/2}$)
- From $CC_{1/2}$, we can analytically estimate **CC of the merged dataset against the true** (usually unmeasurable) **intensities** using

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

- (Karplus and Diederichs (2012) *Science* **336**, 1030)

Data CCs



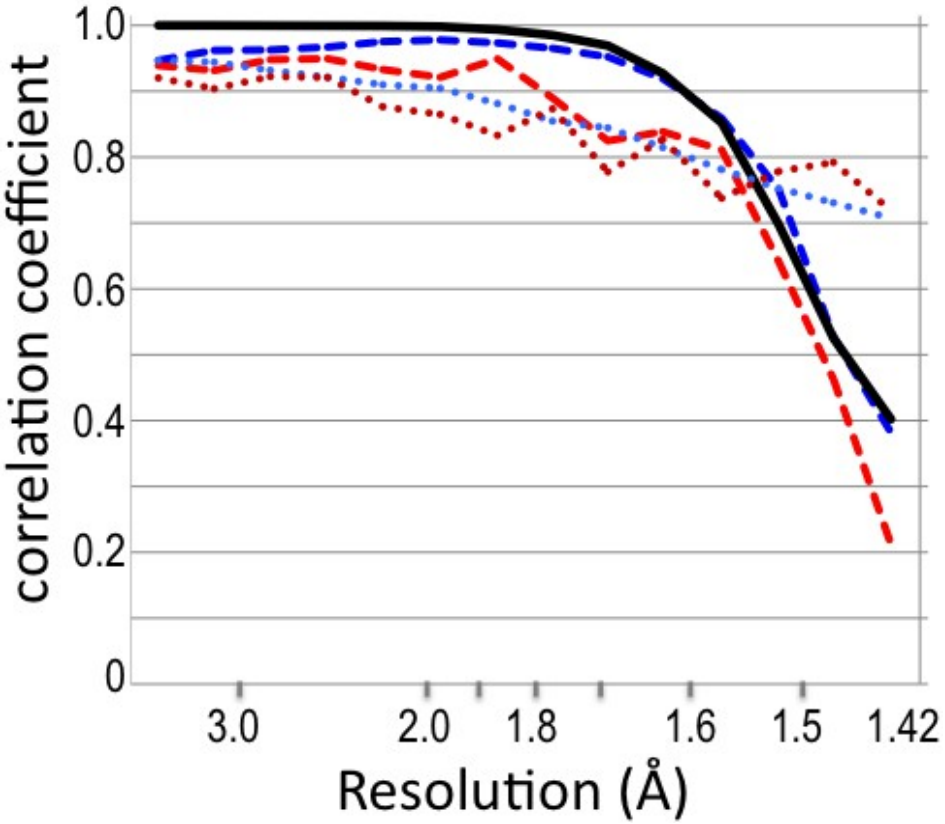
$CC_{1/2}$ ◊

CC* Δ

I/sigma

Model CCs

- We can define CC_{work}, CC_{free} as CCs calculated on F_{calc}^2 of the working and free set, against the experimental data
- CC_{work} and CC_{free} can be directly compared with CC*



— CC*

Dashes: CC_{work}, CC_{free} against weak exp'tl data

Dots: CC'_{work}, CC'_{free} against strong 3ELN data

Four new concepts for improving crystallographic procedures

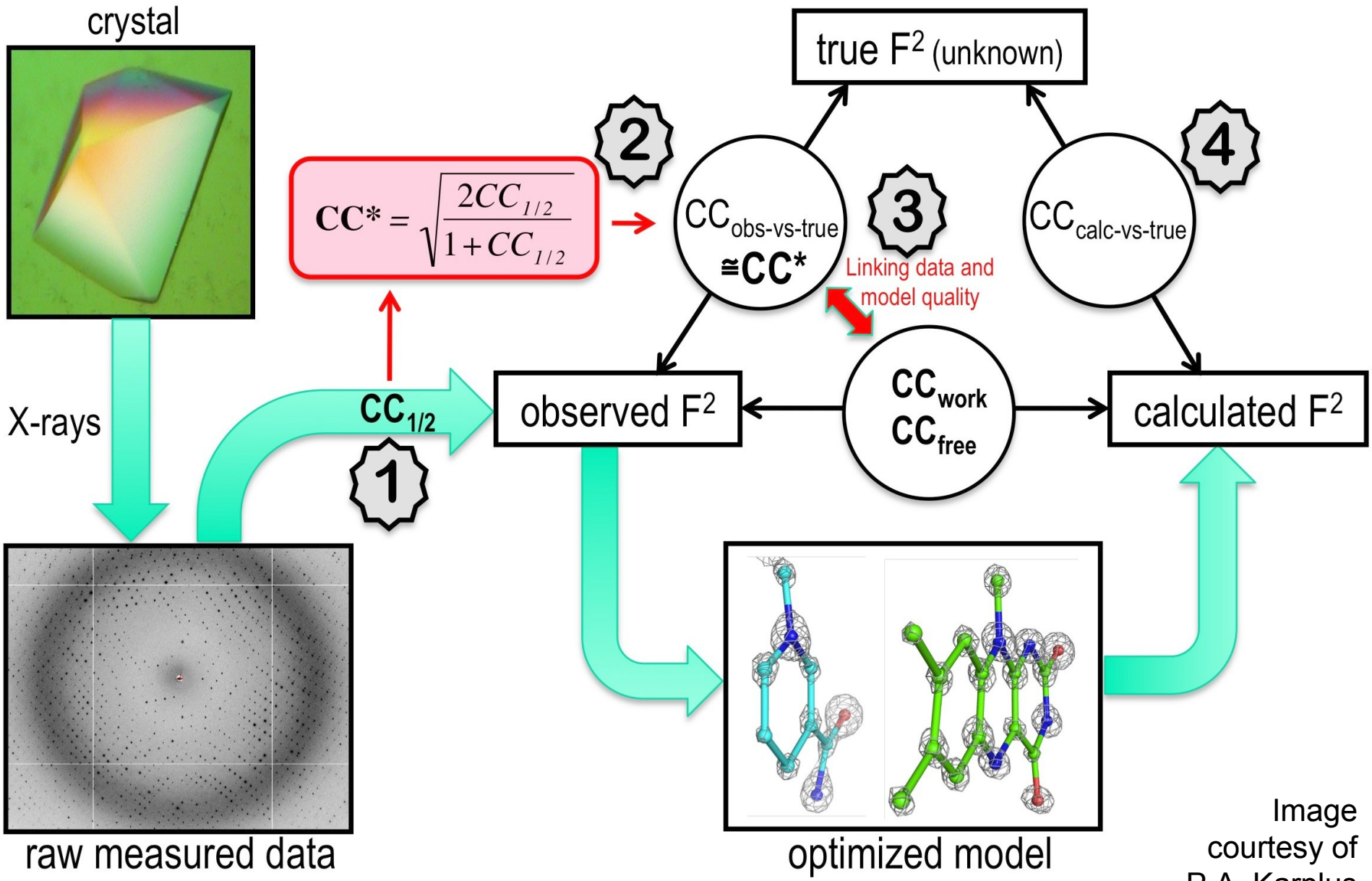


Image courtesy of P.A. Karplus

Summary

- To predict suitability of data for downstream calculations (phasing, MR, refinement), we should use indicators of merged data precision
- R_{merge} should no longer be considered as useful for deciding e.g. on a high-resolution cutoff, or on which datasets to merge, or how large total rotation
- I/σ has two drawbacks: programs do not agree on σ, and its value can only rise with multiplicity
- $CC_{1/2}$ is well understood, reproducible, and directly links to model quality indicators

References

P.A. Karplus and K. Diederichs (2012) Linking Crystallographic Data with Model Quality. *Science* **336**, 1030-1033. see also: P.R. Evans (2012) Resolving Some Old Problems in Protein Crystallography. *Science* **336**, 986-987.

K. Diederichs and P.A. Karplus (2013) Better models by discarding data? *Acta Cryst.* D**69**, 1215-1222.

P. R. Evans and G. N. Murshudov (2013) How good are my data and what is the resolution? *Acta Cryst.* D**69**, 1204-1214.

Z. Luo, K. Rajashankar and Z. Dauter (2014) Weak data do not make a free lunch, only a cheap meal. *Acta Cryst.* D**70**, 253-260 .

J . Wang and R. A. Wing (2014) Diamonds in the rough: a strong case for the inclusion of weak-intensity X-ray diffraction data. *Acta Cryst.* D**70**, 1491-1497.

Diederichs K, "Crystallographic data and model quality" in Nucleic Acids Crystallography. (Ed. E Ennifar), Methods in Molecular Biology (in press).

Thank you!

PDF available – pls send email to
kay.diederichs@uni-konstanz.de