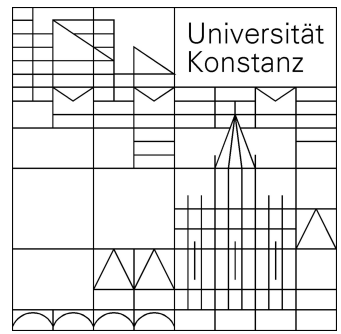
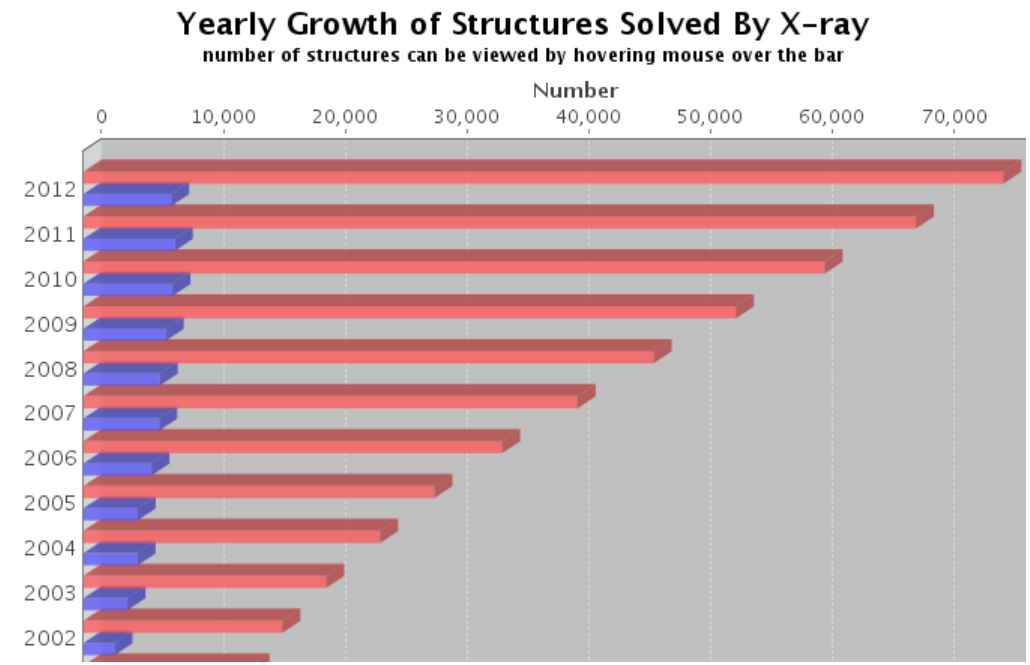
Linking crystallographic model and data quality: theory and applications

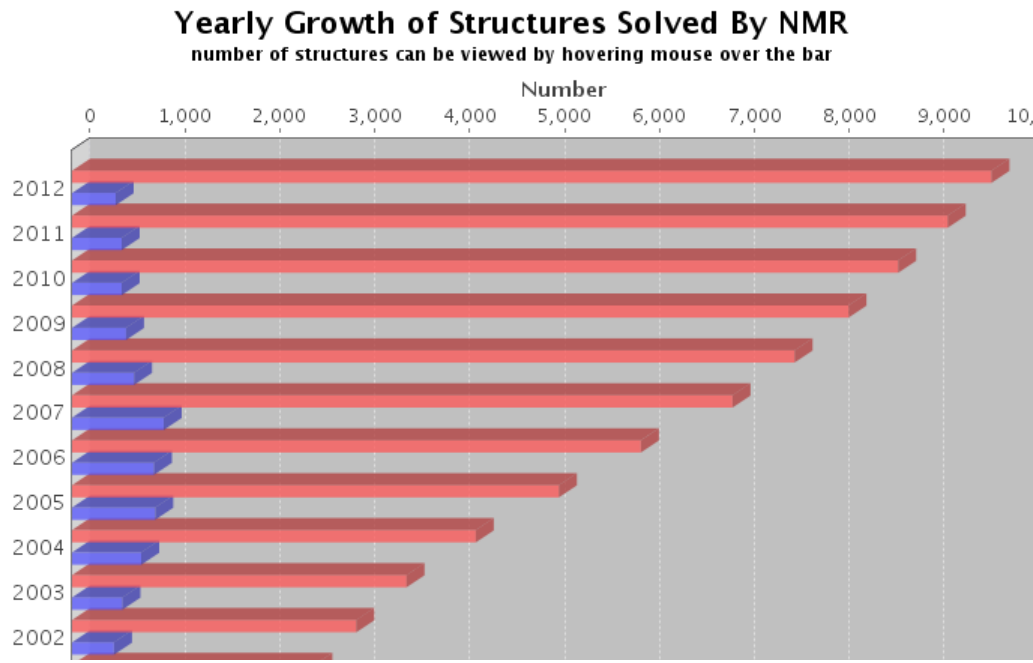
Kay Diederichs



University of Konstanz

Crystallography is highly successful





Can we do better?

Errors in experimental data

Error = *random* + *systematic*

Random = counting + detector

Systematic = Radiation damage
+ absorption + non-linearities
+ vibrations + instabilities + …

Multiplicity of n reduces the random error by √n

Multiplicity *may* reduce the systematic error by √n, but not necessarily!

"If you don't have good data, then you have no data at all." -Sung-Hou Kim

"If you don't have good data, then you must learn statistics." - James Holton

Crystallographic statistics - which indicators are being used?

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

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

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

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

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

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

- I/σ (for *unmerged* or *merged* data !)

Precision of unmerged and merged data

Precision can be calculated for …
either the unmerged (individual) observations:

$$R_{merge}, R_{meas}, I/\sigma_{obs}$$

… or for the merged data:

$$R_{pim}, CC_{1/2}, I/\sigma_{merged}$$

It is essential to understand the difference, but it is not in the papers or textbooks!

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 !)

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}!

Basic questions: what to optimize? Is the data/model R-value a good predictor/indicator of model quality? How good need the data be; to what extent do the data influence the refinement result? What to refine, and when to stop? Overfitting?

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

Conflicting views

„An appropriate choice of resolution cutoff is difficult and sometimes seems to be performed mainly to satisfy referees ... Ideally, we would determine the point at which adding the next shell of data is not adding any statistically significant information ... R_{merge} is not a very useful criterion.“

P. Evans (2011) An introduction to data reduction: space-group determination, scaling and intensity statistics. *Acta Cryst.* **D67**, 282

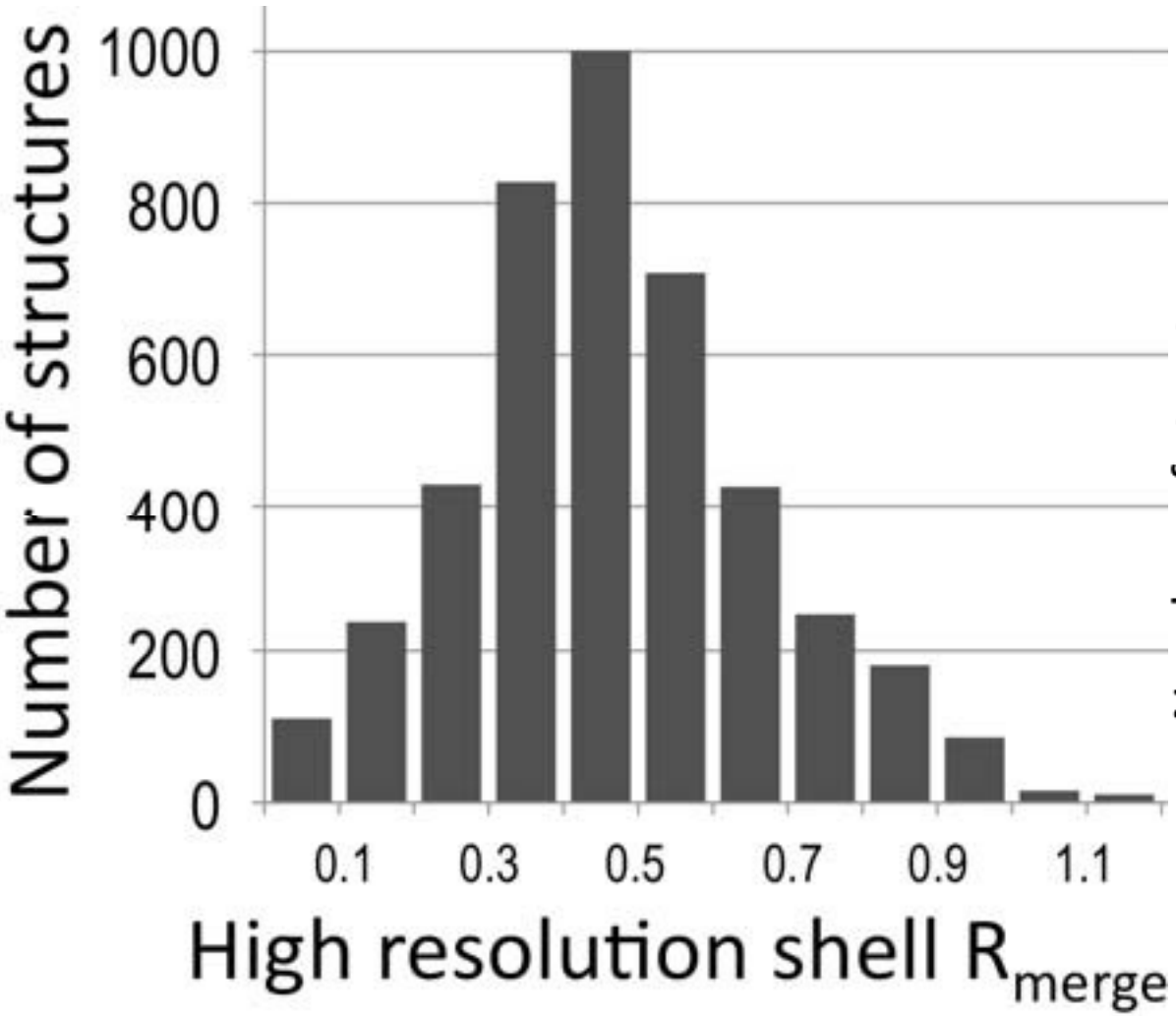
"At the highest resolution shell, the R_{merge} can be allowed to reach 30–40% for low-symmetry crystals and up to 60% for high-symmetry crystals, since in the latter case the redundancy is usually higher."

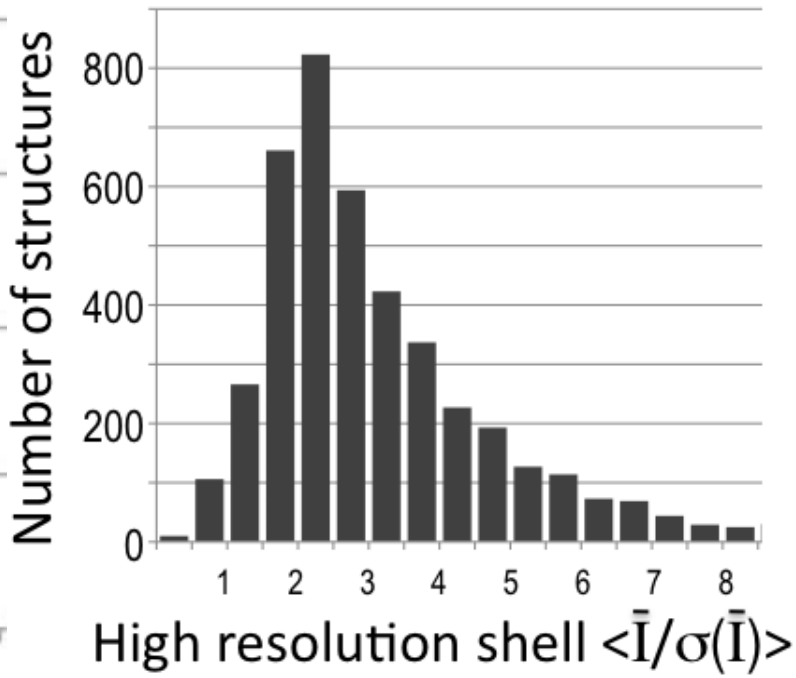
A. Wlodawer, W. Minor, Z. Dauter and M. Jaskolski (2008) Protein crystallography for non-crystallographers, or how to get the best (but not more) from published macromolecular structures. *FEBS J.* **275**, 1

“... the accepted resolution limit is where the I/sigI falls below about 2.0. R_{merge} may then reach 20-40%, depending on the symmetry and redundancy.“

Z. Dauter (1999) Data-collection strategies. *Acta Cryst* **D55**, 1703

2010 PDB depositions





The asymptotic behaviour of model and data R-values is different at high resolution

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

Data R-values (R_{pim} , R_{merge} , R_{meas}): with resolution, go to *infinity* since the numerator is *constant* (determined by background), and the denominator approaches *zero*

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

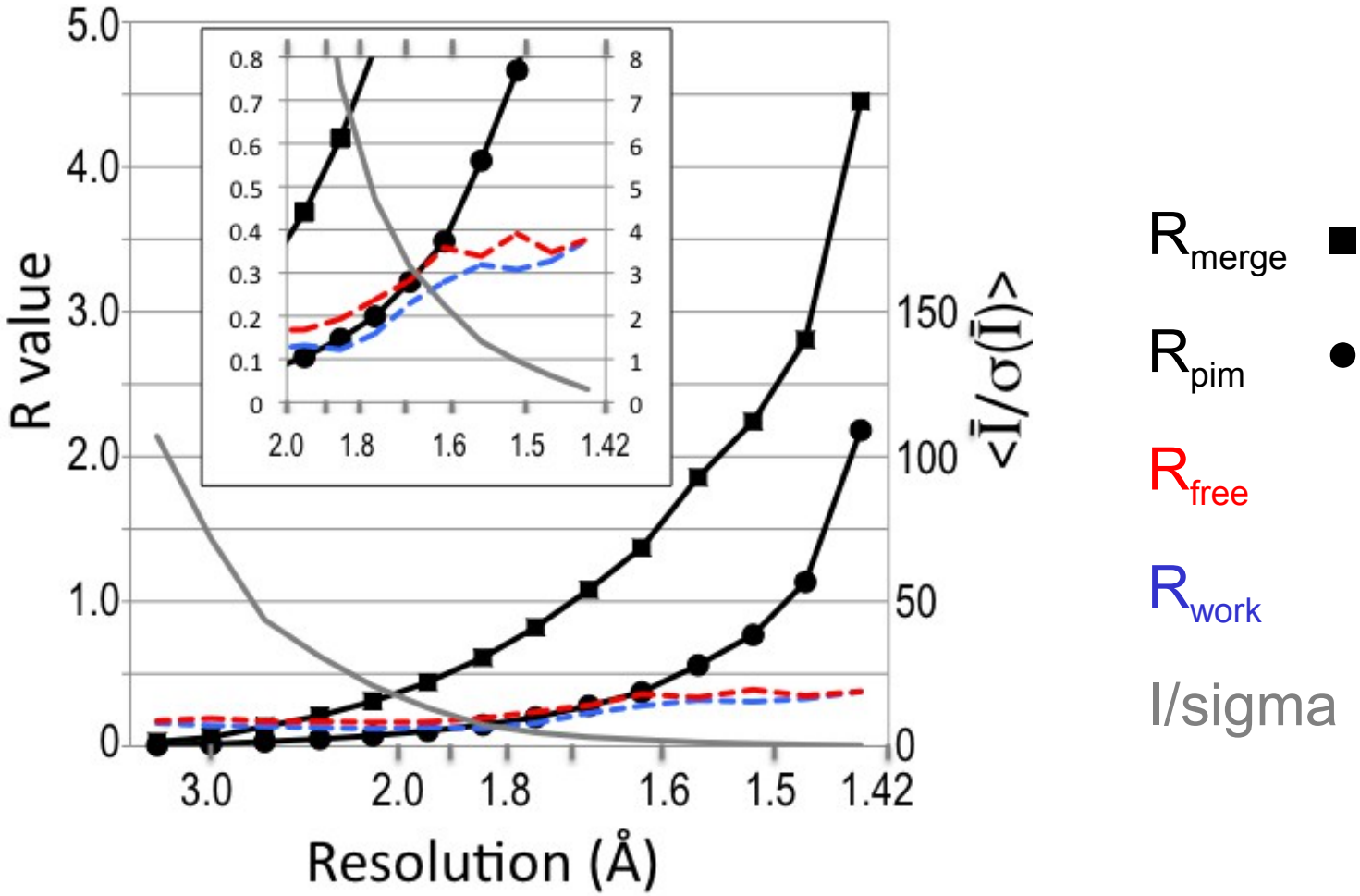
Model R-values ($R_{work/free}$): the ratio of numerator and denominator approaches a *constant* for a random or wrong model (Wilson 1950), or for random data (Evans & Murshudov 2013)

This means that at high resolution, a quantitative relation cannot exist between model and data R-values !

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.

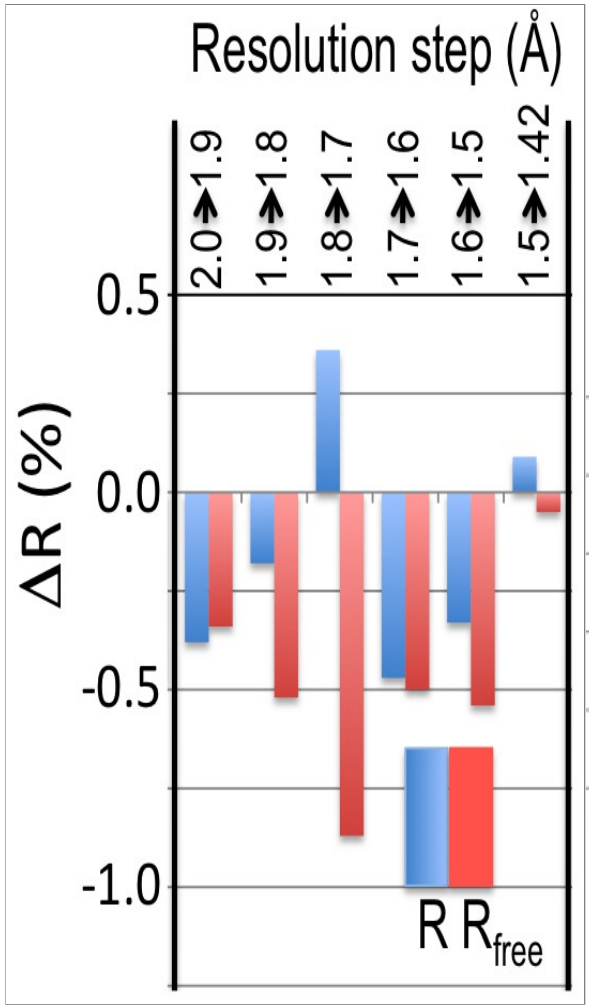
Example: Cysteine DiOxygenase (CDO; PDB 3ELN) re-refined against 15-fold weaker data

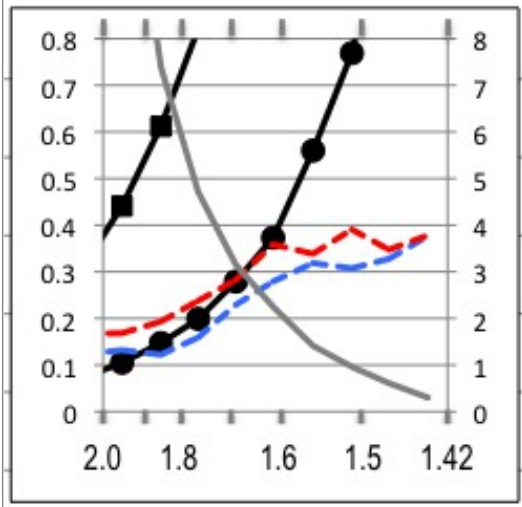


Is there information beyond the conservative hi-res cutoff?

"**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`)
- $\Delta R = R_{1.9}(2.0) - R_{2.0}(2.0)$

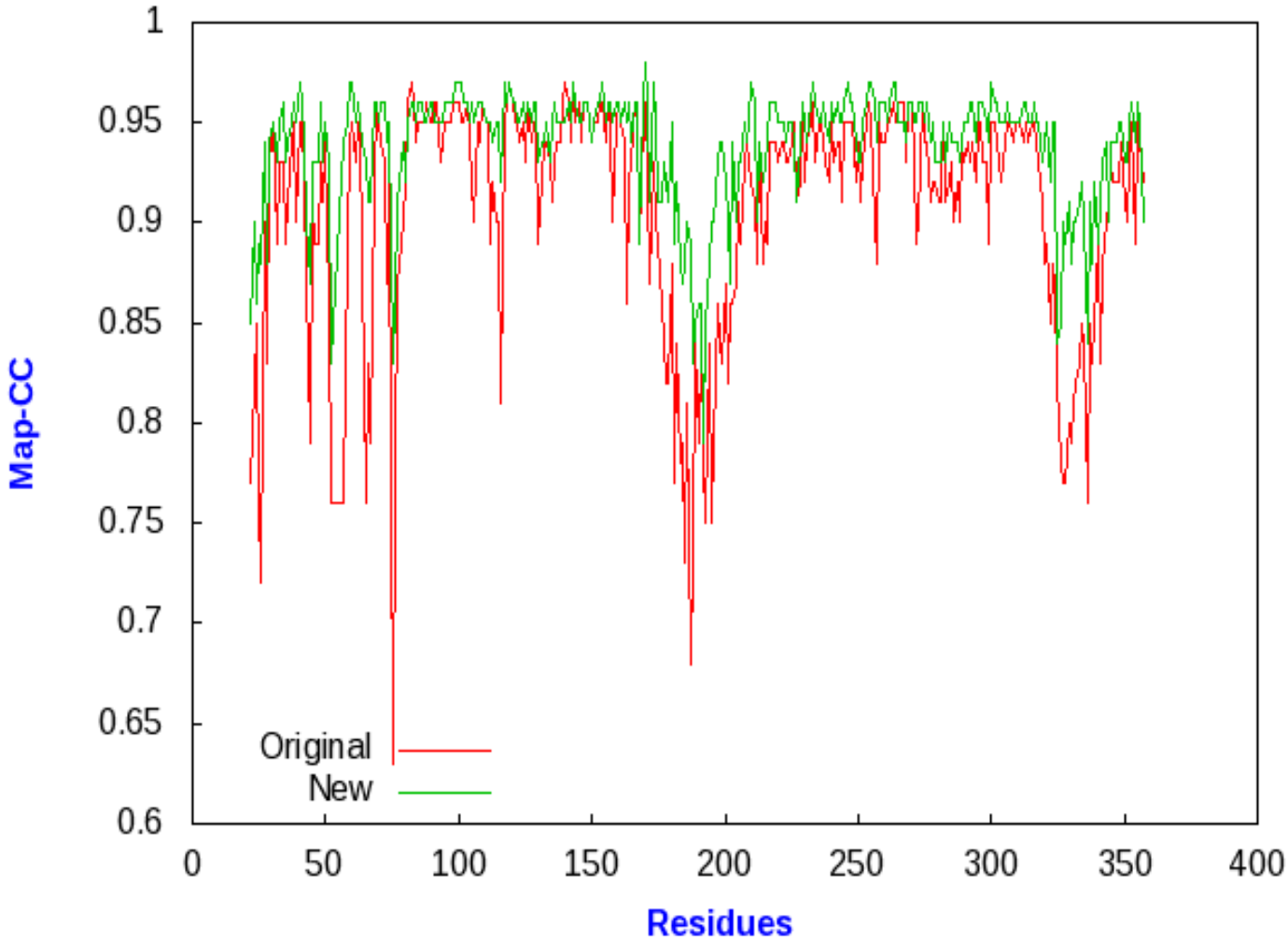




Do the maps really get better?

- cooperation with JCSG: Henry van den Bedem, Ashley M. Deacon, Abhinav Kumar
- identified 5 structures where re-processing of existing raw data permits to extend the resolution by another 0.2-0.3Å, with full completeness
- reprocessing with XDS/MOSFLM; refinement with refmac/phenix.refine
- calculate real-space CC of map and model, using EDSTATS (CCP4) or phenix.resolve

Example 1/5


Map-CC
1
0.95
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0
50
100
150
200
250
300
350
400
Residues
Original
New

Summary I

- R_{merge} should no longer be considered as useful for deciding on a high-resolution cutoff
- The *paired refinement technique* can prove that data should be used to higher resolution than a (R_{merge}-based) conservative cutoff suggests

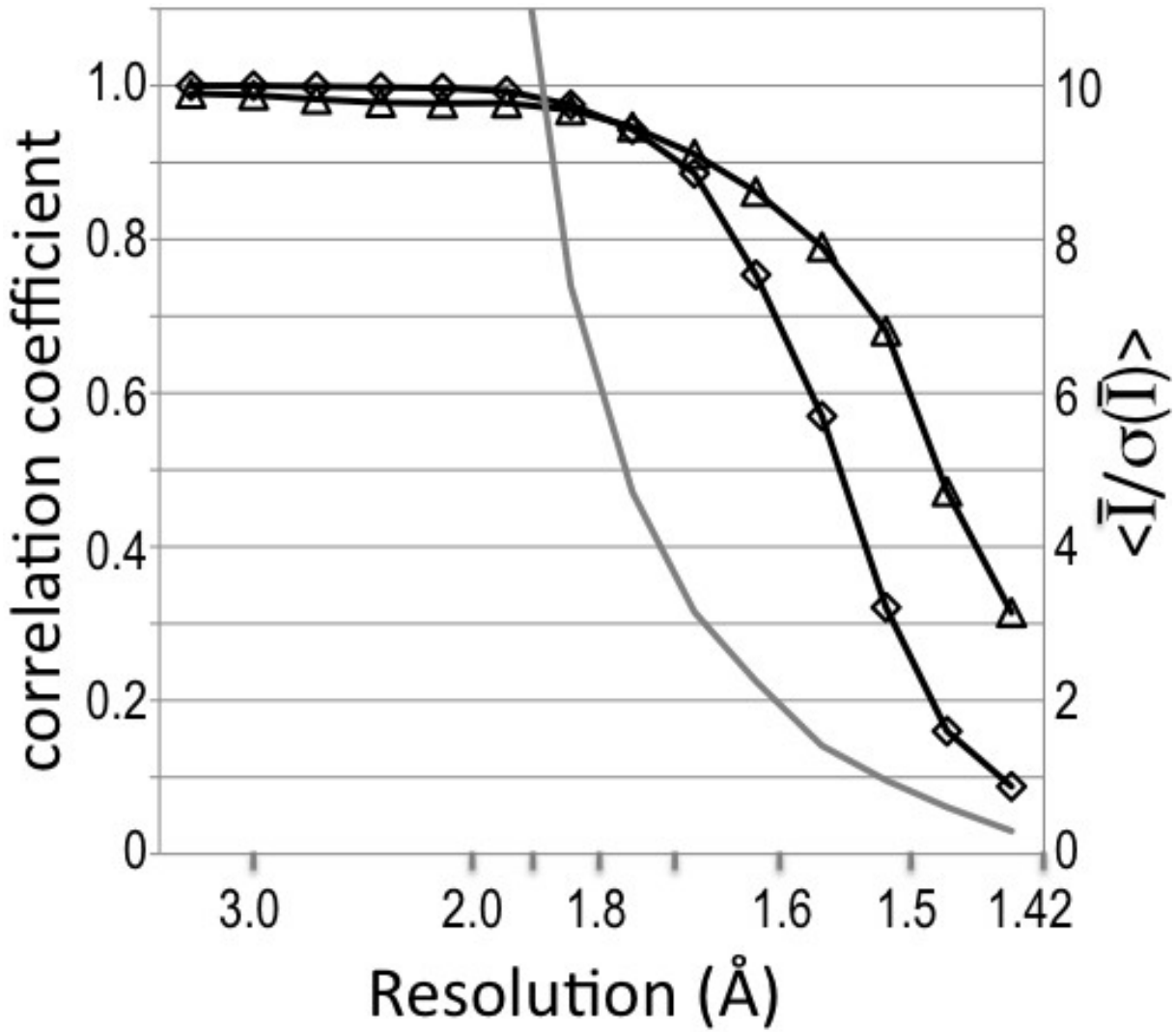
measuring data quality with a correlation coefficient

- 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 full 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



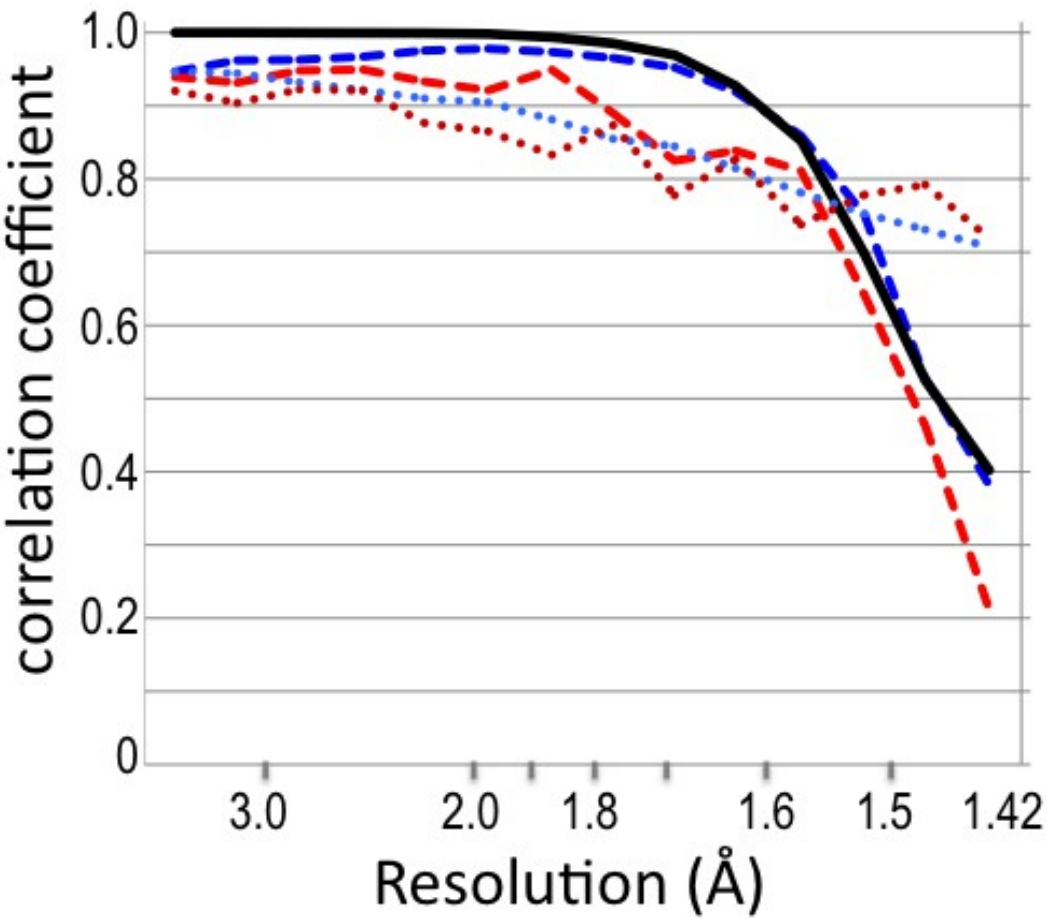
$CC_{1/2}$ ◊

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

Quantitative relation between data and model CCs

- Refinement should make CC_{work} converge towards CC* (from lower values)
- Inadequate model, or wrong space group, or systematic errors in data processing: CC_{work} remains < CC*
- If CC_{work} > CC*: the model is closer to the data, than the truth is to the data : “overfitting”

Summary II

- $CC_{1/2}$ assesses the statistical significance of data
- $CC^* = \sqrt{\frac{2\,CC_{1/2}}{1 + CC_{1/2}}}$ tells us the agreement between experimental data and true data (!)
- CC* is the **upper limit** for the CC_{work} /CC_{free} model quality indicators
- $CC_{1/2}$, CC*, CC_{work}/CC_{free} table can be obtained from 1.8.2 Phenix distribution; the routine is called *phenix.cc_star*

Four new concepts for improving crystallographic procedures

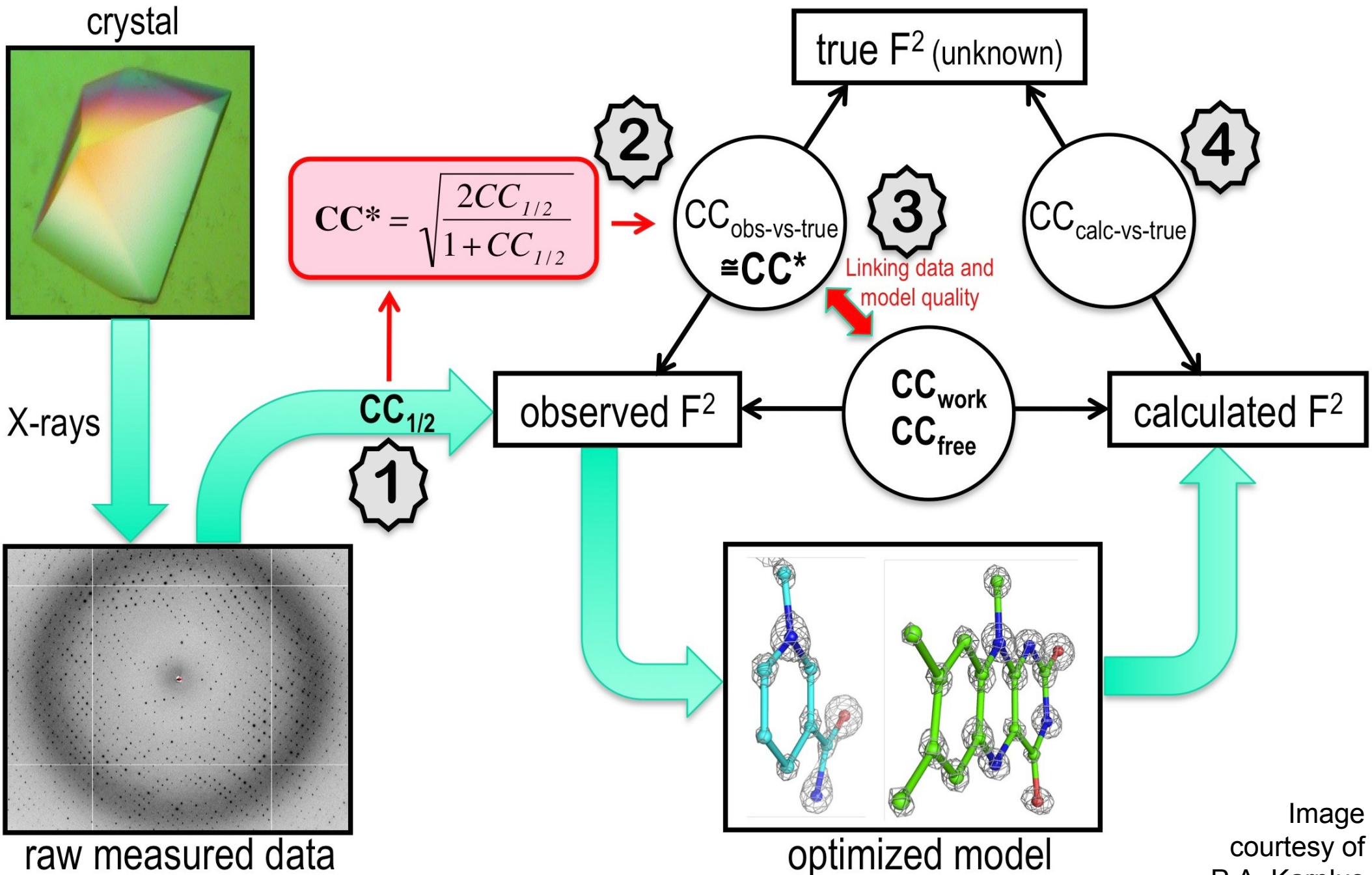


Image courtesy of P.A. Karplus

Acknowledgement

Andy Karplus,
Oregon State
University
(Corvallis, OR)

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.