GΦL

Global Phasing Limited

The raw, the cooked and the medium-rare

Unmerged diffraction data as a rich source of opportunities for data re-use and improvements in methods and results

G. Bricogne, C. Flensburg, R.H. Fogh, P.A. Keller,
I.J. Tickle, C. Vonrhein

Global Phasing Ltd, Cambridge, UK

IUCr2023 CommDat Workshop
Raw Diffraction Data Reuse: the Good, the Bad and the Challenging
Melbourne, 22 August 2023

GΦL
Global Phasing Limited

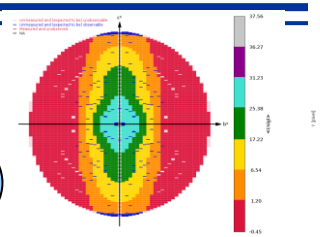
Who we are …

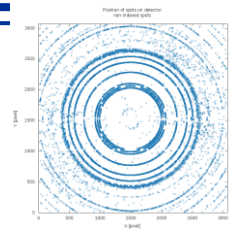
- Created in 1997 in response to industrial interest (since 1994) in using forerunning academic work on the use of Bayesian statistics for experimental phasing (SHARP) and structure refinement (BUSTER)
- That interest followed from the need to robustify, accelerate and automate crystal structure determination and refinement for the purposes of SBDD - without this, structural results typically arrived too late.
- A timely word of advice from Tom Blundell (in the mid-1990s): "*If you are going to work with industry, you absolutely have to be polygamous*"
- Consortium model, to **do science rather than make money** – still free from any borrowing after 26 years of operation
- Now comprises 40 members, combining Big Pharma and smaller drug discovery companies, together with CROs (big and small) servicing them – so we have a complete eco-system of the MX-based SBDD world
- We sit at a triple point between Academia, Industry and Synchrotrons, with a mission to improve performance, automation and integration in the use of MX for structure-based drug discovery and structural biology in general

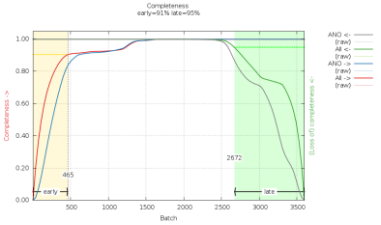
Global Phasing Limited

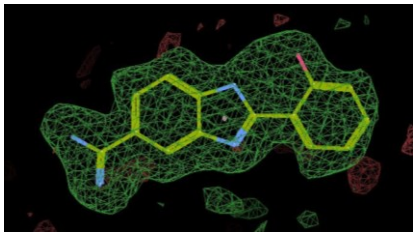
… and what we do

- Data collection and experiment
  - Global Phasing **workflow** (MXCuBE, GDA, ISPyB)
  - **Multi-sweep strategies** for native and phasing data collection
- Data processing
  - **autoPROC** (XDS, CCP4, with much “added value”)
  - **STARANISO** (anisotropy characterisation and mitigation)
  - multi-lattice, ice rings, pruning poor image ranges, bad pixels
- Ligands
  - **BUSTER** (ligand detection maps)
  - **Grade** restraints dictionary generation (CSD, QM)
  - **RhoFit** (automated fitting into density)
- Refinement
  - **BUSTER** (structure refinement)
  - **Pipedream** (ligand detection & refinement pipeline)
- Validation, reporting and deposition
  - **buster-report, CRIMS-Pipedream interface**
  - **PDBx/mmCIF** output combining model and rich data
  - Advanced methods for data validation and “auditing”









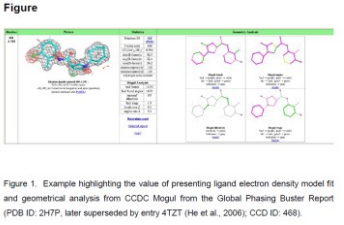


Figure 1. Example highlighting the value of presenting ligand electron density model fit and geometrical analysis from CCDC Mogul from the Global Phasing Buster Report (PDB ID: 2H7P, later superseded by entry 4TZT (He et al., 2006); CCD ID: 468).

Global Phasing Limited

Why we are interested in raw data reuse

- We are application developers, not a synchrotron, but one of our applications (autoPROC) is widely used at synchrotrons and by synchrotron users.
- Enabling autoPROC to process raw diffraction images autonomously (i.e. with minimal local wrappers) and in a maintainable manner has required tackling head-on the bewildering diversity of instrumental configurations and image formats over the past two decades by maintaining an up-to-date record of the information not available in image headers nor master files in a publicly accessible Wiki page (http://www.globalphasing.com/autoproc/wiki/index.cgi?BeamlineSettings) now replaced by a text file (detector_site.def) included in our distribution, containing detector type with serial number, date limits, header convention for beam-centre coordinates, sense of Omega axis rotation, list of known bad pixels, and more.
- Much of our user support work involves the diagnosis and remediation of problems that provide a steady stream of incentives to (1) keep improving the software and (2) validate these improvements by (re-)processing other datasets than those against which they were initially developed.
- This has constantly required trying to **F**ind, **A**ccess and **R**e-use datasets of raw diffraction images from diverse repositories and has confronted us with a great lack of **I**nteroperability between these repositories (more on this in MS A118).

Global Phasing Limited

Diversion: a glimpse at detector_site.def

22 Synchrotrons

ALBA (http://www.cells.es/)
ALS (http://www-als.lbl.gov/)
APS (http://aps.anl.gov/)
Australian Synchrotron (http://www.synchrotron.org.au/)
BESSY (http://www.helmholtz-berlin.de/bessy-mx)
CHESS (https://www.chess.cornell.edu/)
CLSI (http://cmcf.lightsource.ca/)
Diamond (http://doc.diamond.ac.uk/MXManual/analysis/detector.html)
ELETTRA (https://www.elettra.eu/)
ESRF (http://www.esrf.eu/)
LNLS https://www.lnls.cnpem.br/
MAX-IV (https://www.maxiv.lu.se/)
NSLS (http://www.px.nsls.bnl.gov/)
NSLS-II (https://www.bnl.gov/ps/)
PAL/PLS (https://pal.postech.ac.kr/paleng)
PETRA-III (http://petra3.desy.de/)
Photon Factory (http://pfwww.kek.jp/)
SLS (http://www.psi.ch/sls/swiss-light-source)
SOLEIL (http://www.synchrotron-soleil.fr/)
SPring-8 (http://www.spring8.or.jp/en/)
SSRF (http://www.sinap.ac.cn/e-ssrf/)
SSRL (http://ssrl.slac.stanford.edu/)

76 Beamlines

124 detectors
- 1 CCD
- 1 MAR
- 1 MAR345
- 1 RAYONIX
- 1 RIGAKU/Pilatus
- 5 RIGAKU
- 15 HDF5/Eiger
- 15 MARCCD
- 28 ADSC
- 56 Pilatus

224 different variants over time

Two beamlines (ALS 5.0.2 and BESSY 14.2) have had **5** different detectors, and two others (ALS 5.0.1 and SLS PX-I) **4** different detectors, over the period tracked (2002-2023).

The detector with the greatest number of variants over time (since 20 March 2010) is the Pilatus 6M on SSRL 12-2 (S/N 60-0101), with **7** variants.

Global Phasing Limited

How we re-use raw data and why

- We view raw data archiving as serving two different purposes:
  - Recording the direct experimental evidence supporting structural results in the PDB
  - Feeding the “virtuous circle” of mutual iterative improvements between software and final results that is empowered by the availability of “upstream data”, as happened with the improvement of refinement programs thanks to the deposition of merged diffraction data.
- Our primary interest in raw data springs from that *second* purpose, as part of our efforts to “push the frontiers” along the whole sequence of steps in MX, from diffraction experiment to deposition of final results into the PDB.
- This has two consequences:
  - As we typically want to follow the cascade of improvements starting with raw data reprocessing and ending with maps, difference maps, models and validation results, we have tended to concentrate on **accessing raw data via the DOIs given in PDB entries** in order to move upstream of the deposited (merged) data.
  - Deciding on what frontiers most need pushing naturally leads to the question: **what kinds of upstream data** are necessary to test different categories of ideas and implementations, then validate them?
- Our viewpoints on the whole topic of raw data archiving are thus based on our practical experience with their re-use in the context of these activities - excluding politics and polemics (but see MS A118) as well as large-scale efforts that have not yet produced software that facilitates our work.

GΦL
Global Phasing Limited

How “upstream” for which developments?

Requiring access to raw diffraction images

- Spot collection
- Indexing
- Determination of unit cell and crystal orientation
- Reflection profile estimation
- Integration from profile estimates
- Exclusion of bad image ranges, and bad ranges of poor images
- Remediation of dynamic shadows
- LP corrections
- Outlier rejection and production of a first cut of internally scaled and unmerged intensities

Feasible from ordinary scaled, unmerged intensities

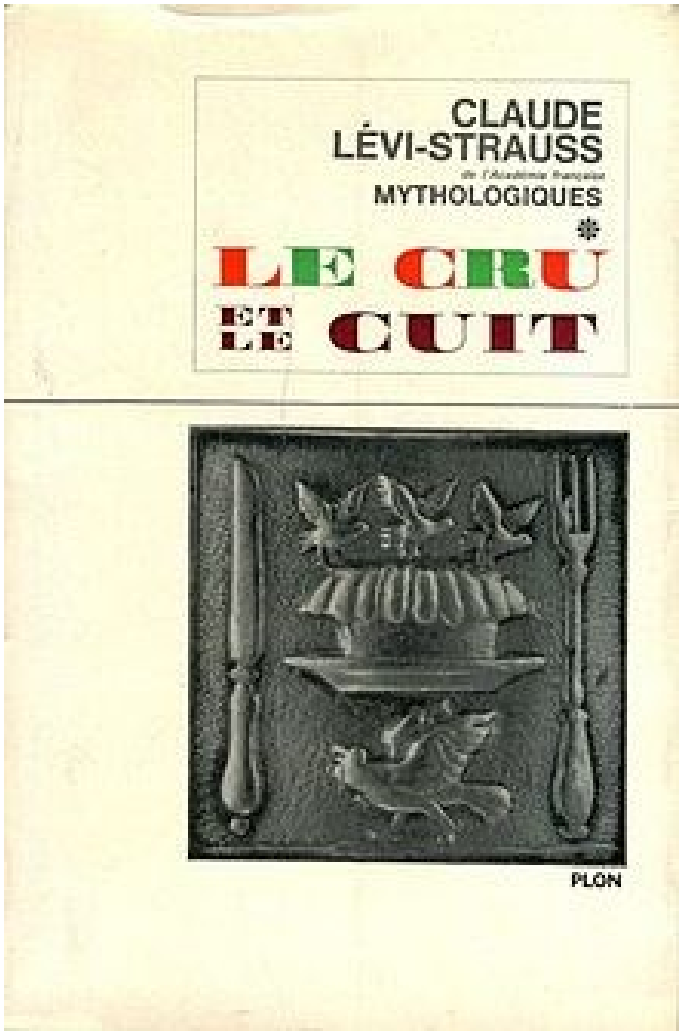
- Radiation damage detection and characterisation.
- Exploration of improved scaling models (e.g. beyond the image scale factor plus Biso level)

A natural question arises:

- *What is missing from “ordinary” scaled and unmerged intensities that limits their extra usefulness, requiring a return to raw images?*
- *Could this information be added in the form of extra reflection data that could “enrich” PDB files and thus be archived by deposition?*

GΦL
Global Phasing Limited

Raw, Cooked, and Medium-rare



In Levi-Strauss's Book

Raw: as occurring in Nature
Cooked: as transformed by cultural processes of increasing sophistication

In MX data management:

Raw = diffraction images
Cooking = processing
Cooked = merged reflection data against which model refinement is performed
Medium-rare: enriched unmerged reflection data, retaining some of the "juices" (metadata) of the raw data

GΦL
Global Phasing Limited

The scientific case for the deposition of *enriched unmerged* data into the PDB (1)

- The idea is to **enrich the datablocks in mmCIF files with reflection-wise metadata** not ordinarily carried over as part of deposited unmerged data, such as image number and detector coordinates at which individual reflections occur, as well as extra instrumental configuration metadata
- This information is **readily available from processing programs** for data collected by the rotation method but does not find its way into the data section of the mmCIF files intended for deposition
- Making it possible to enrich these files with that extra information through suitable extensions of the mmCIF dictionary has been the goal of the *Subgroup on Data Collection and Processing of the PDBx/mmCIF Working Group of the wwPDB* that has been active since October 2020
- This is achievable with **modest storage requirements** compared to the archival of raw images, while already creating a **standardised** resource capable of supporting many of the improvements in scaling and merging methods (and results) that are currently only possible from raw images
- A key benefit of this approach is that it circumvents the patchiness and un-FAIRness of the current status of Raw Data archives.

GΦL
Global Phasing Limited

The scientific case for the deposition of *enriched unmerged* data into the PDB (2)

- Examples of possibilities for improving initially performed scaling/merging steps and for extraction of further data:
  1. production of full validated data quality metrics that are often incomplete or inconsistent in deposited merged data;
  2. detection of problematic images and image ranges, and remediation by their selective exclusion from scaling/merging;
  3. anisotropic diffraction limit analysis (or re-analysis) with STARANISO, if not already performed;
  4. extraction of previously unexploited anomalous signal and computation of anomalous difference Fourier maps;
  5. **"reflection auditing"** by tracing outliers detected at the refinement stage back to their unmerged contributors in terms of specific image numbers and detector positions, thus diagnosing ice rings, poor beamstop masks, angular overlaps, etc. ;
  6. **detection of radiation damage** via Fearly - Flate maps; adapting parametrisation to patterns of structural radiation damage.
- The extensions under active development also enable the archiving of unmerged serial (SSX/SFX) data (Aaron Brewster's contribution)
- "Extending the mmCIF dictionary with so many new items? You must be dreaming …" *No! It can be done, and it has already been done in the previous phase of activity of this SubGroup!*

GΦL

Global Phasing Limited

PDBx/mmCIF Home Dictionaries Documentation Downloads Contact Us

Dictionary Revision History

Announced on 30 March 2021

Over **50** new items were **collaboratively added** to the mmCIF dictionary in March 2021 to allow the archiving of

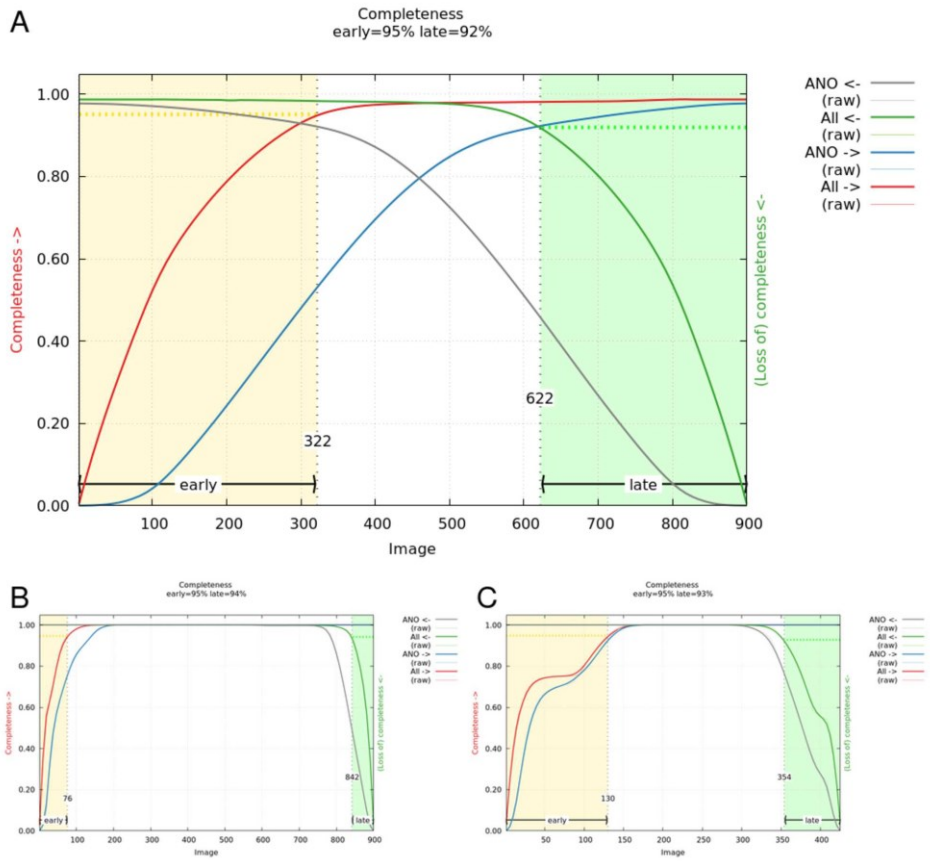
- **anisotropic** data statistics
- **unmerged** refection data
- missing quality metrics for **anomalous** diffraction data

```
5.339   2021-02-16

                    Changes (Global Phasing Ltd.):
                    + Add items for anomalous diffraction statistics
                       _reflns.pdbx_redundancy_anomalous, _reflns.pdbx_CC_half_anomalous,
                       _reflns.pdbx_absDiff_over_sigma_anomalous, _reflns.pdbx_percent_possible_anomalous,
                       _reflns_shell.pdbx_redundancy_anomalous, _reflns_shell.pdbx_CC_half_anomalous,
                       _reflns_shell.pdbx_absDiff_over_sigma_anomalous,
                       _reflns_shell.pdbx_percent_possible_anomalous
                    + Add items to cater for anistropic diffraction
                      Ellipsoid fit to the cut-off surface:
                         _reflns.pdbx_aniso_diffraction_limit_axis_?_ortho[?],
                         _reflns.pdbx_aniso_diffraction_limit_?
                      Anisotropic B tensor:
                         _reflns.pdbx_aniso_B_tensor_eigenvector_?_ortho[?],
                         _reflns.pdbx_aniso_B_tensor_eigenvalue_?
                      Statistics specific to anisotropic diffraction:
                         _reflns.pdbx_percent_possible_*,
                         _reflns_shell.pdbx_percent_possible_*
                         for ellipsoidal/spherical and anomalous/non-anomalous diffraction.

                      Also add _reflns.pdbx_orthogonalization_convention and new subcategories
                      unit_vector and eigendecomposition
                    + Add items for a per-reflection signal and parameter-free definition of
                      the cut-off surface
                      _reflns.pdbx_observed_signal_threshold, _reflns.pdbx_signal_type,
                      _reflns.pdbx_signal_details, reflns.pdbx_signal_software_id,
                      _pdbx_refln_signal_binning.ordinal,
                      _pdbx_refln_signal_binning.upper_threshold,
                      _refln.pdbx_signal, _refln.pdbx_signal_status
                    + Add and modify items to cater better for umerged reflection data
                        Add _diffrn_refln.pdbx_detector_x, _diffrn_refln.pdbx_detector_y,
                           _diffrn_refln.pdbx_scale_value
                        Modify _diffrn_refln.pdbx_image_id, _diffrn_refln.pdbx_scan_angle
```

Global Phasing Limited

Early-minus-late differences

Optimal segmentation into ”early” and ”late” datasets by bilateral cumulative completeness analysis

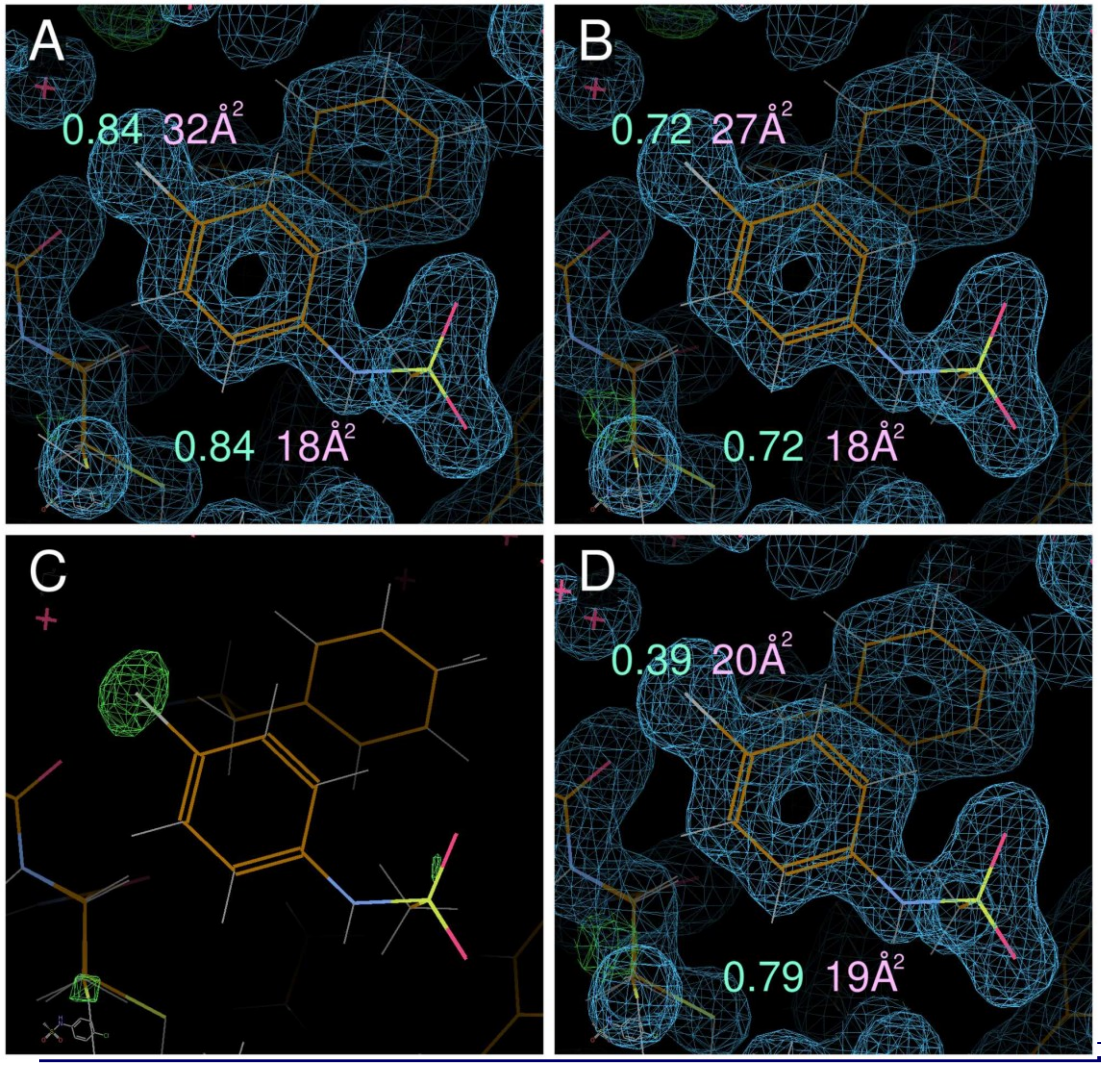


(A) 5SRX (Correy & Fraser, 2022)
(B) 7KDS (Abendroth et al., 2020)
High (cubic) symmetry plus 180-degree rotation range yields a large dose difference
(C) 7WCJ (Sharma et al., 2022)
an unfortunate starting angle for data collection shows as a plateauing of cumulative completeness, leading to a slower increase of cumulative completeness while still accumulating dose

Works from raw diffraction images or ordinary unmerged reflection data

Global Phasing Limited

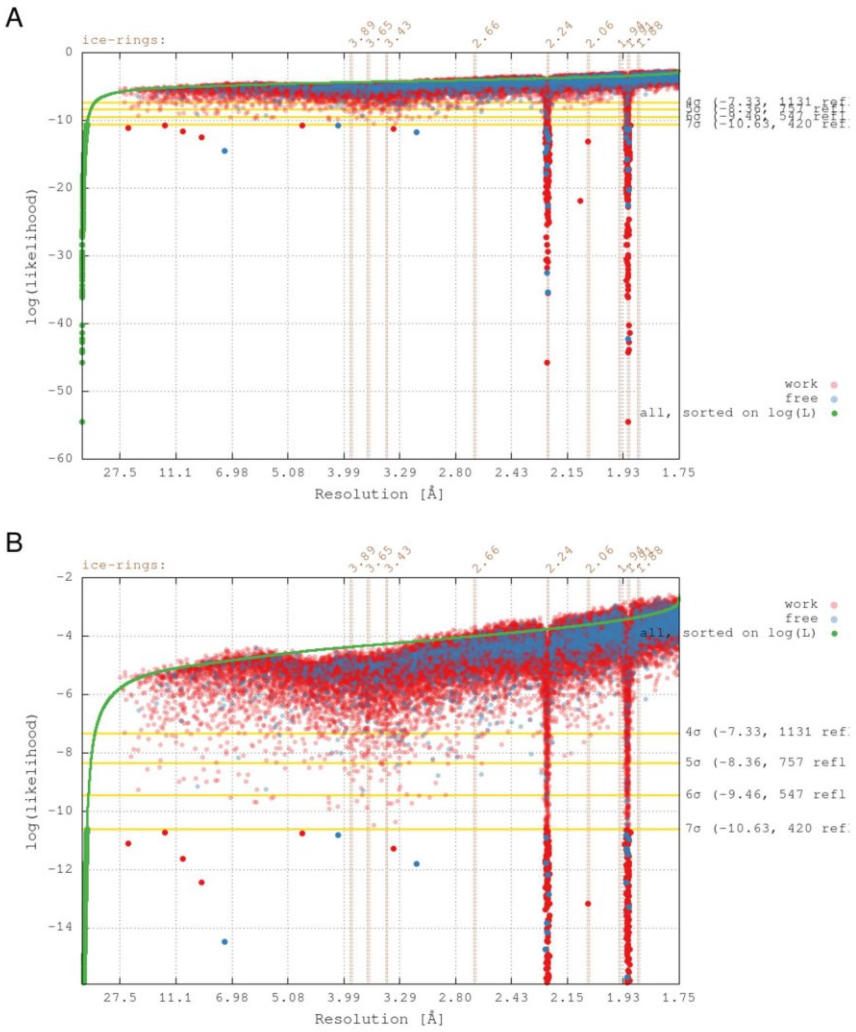
Early-minus late map prompts a change in refinement parametrisation



(A) Deposited model of 6RO in 5KCO: the Cl B-factor stands out
(B) Re-processed raw diffraction data, refining a single occupancy over all compound atoms.
(C) F(early)-F(late) map at 5.0 rms
(D) Re-refinement using separate occupancy parameters for Cl and non-Cl atoms.

Global Phasing Limited

Log-likelihood outliers



Log(likelihood) values of unique reflections after re-refinement of 4Z48 using BUSTER.

(A) Full range of log(likelihood) values.
(B) Close-up to highlight finer details of reflections with smaller log(likelihood) values.

This particular LL-outlier landscape has a plausible explanation, namely an inadequate treatment of ice rings

GΦL
Global Phasing Limited

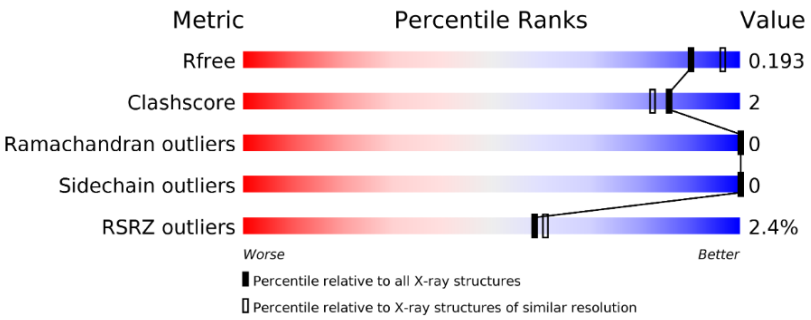
Rationale of "data auditing": LL-outliers should have a "plausible explanation"

- It is bad science to throw away data simply because they do not agree with a model
- If aberrant data are detected through strong disagreement with a model, they should be rejected only if their occurrence can be given a plausible explanation in terms of
  - **sample-related problems** (e.g. cracked crystals, multiple lattices causing overlapped spots)
  - **instrument-related problems** (e.g. bad pixels on the detector surface; modified detection efficiency at module boundaries, faulty modules)
  - **processing problems** (e.g. inadequate beamstop shadow definition, shadowing by goniostat or wires leading to it; inadequate treatment of ice rings or "compound rings"; erroneous rejections of misfits)
- These explanations often are **useful diagnostics of malfunctions**
- For this "auditing" to be possible one must be able to "follow the wires backwards" from merged to unmerged data to images

Global Phasing Limited

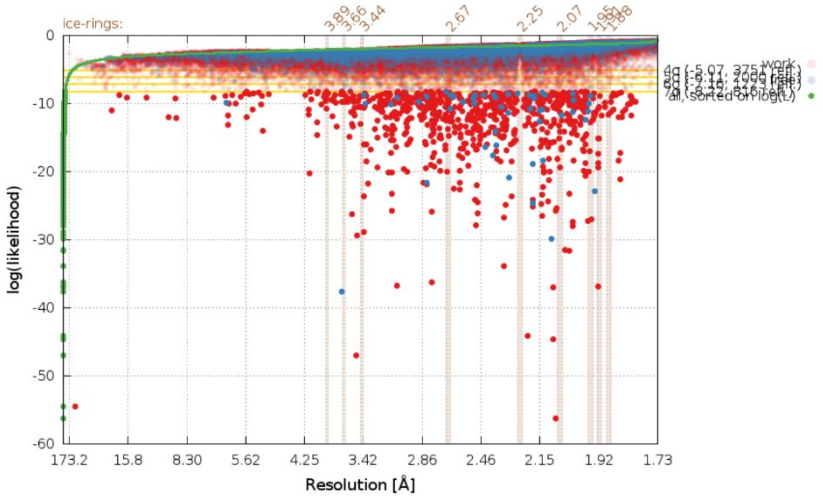
A case study in data auditing: 6vzu

6vzu: a happy-enough looking PDB entry (with raw diffraction images available via SBGrid) …



Summary of Validation Report

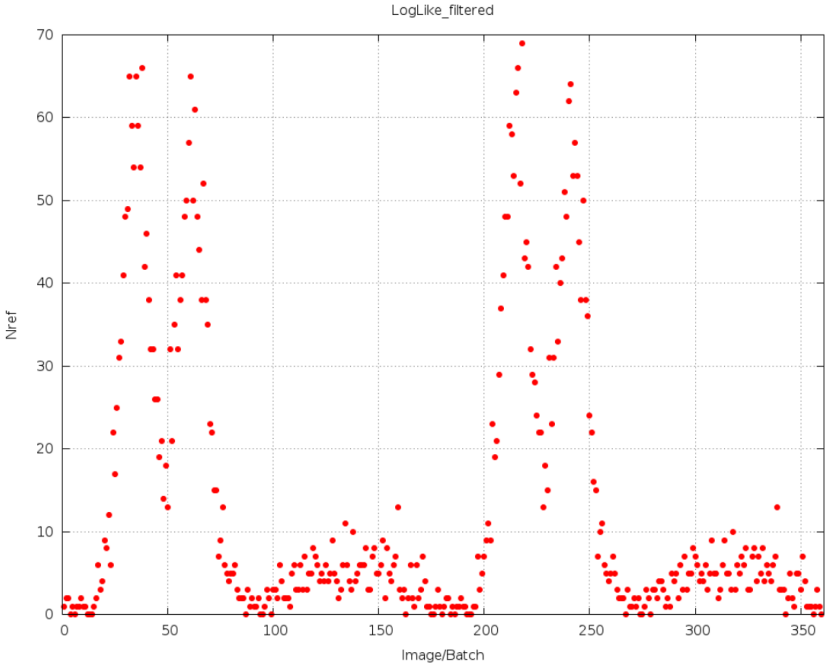
…but with an abnormal abundance and distribution of LL-outliers



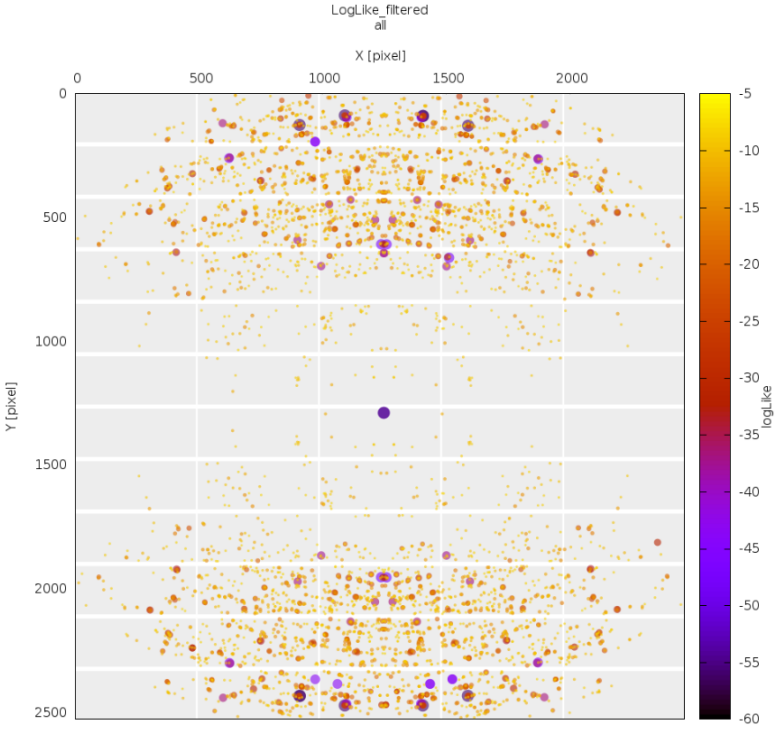
Global Phasing Limited

Mapping the distribution of LL-outliers

Mapped to image number



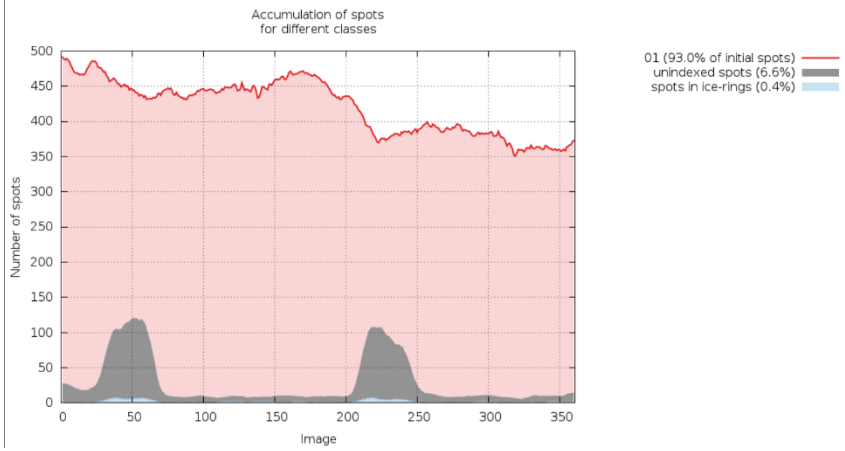
Mapped to detector surface



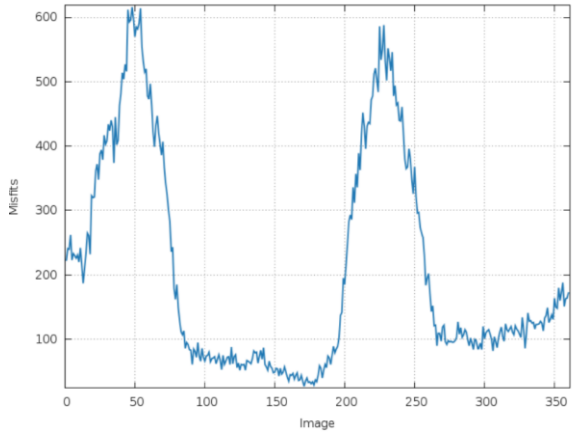
Global Phasing Limited

Reprocess the raw images with autoPROC and look for tell-tale signs

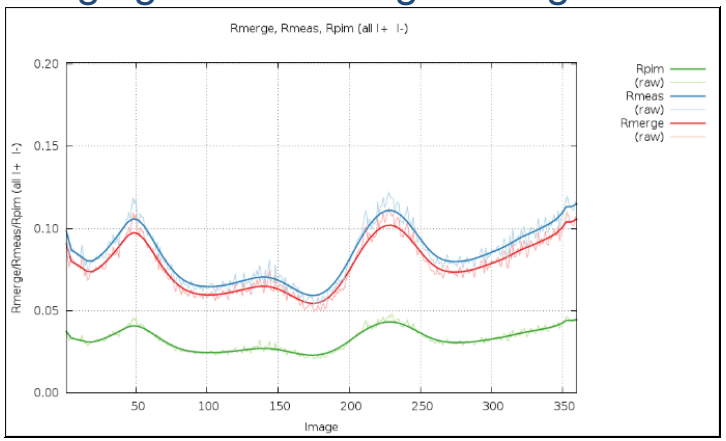
Indexed spots according to image number



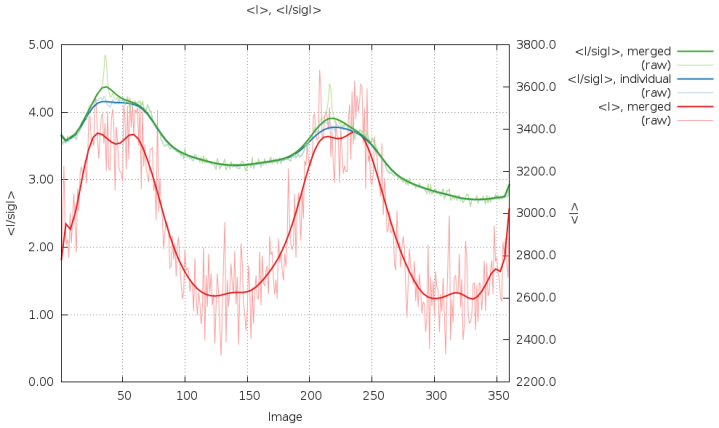
XDS misfits according to image number



Merging Rs according to image number



<I>, <I/sig(I)> according to image number



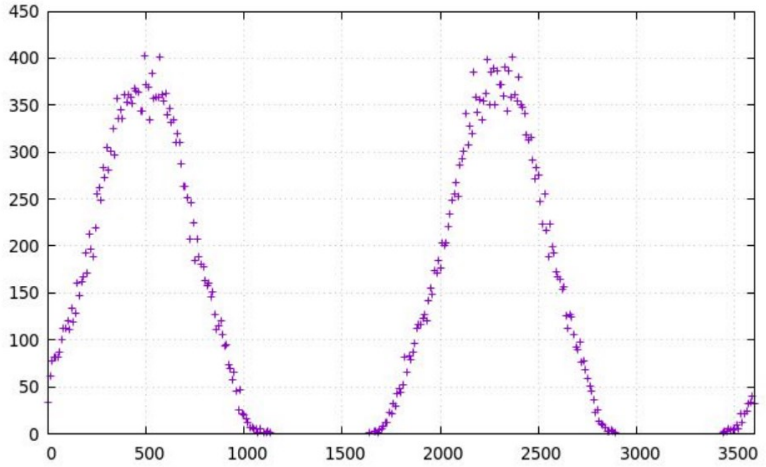
GΦL
Global Phasing Limited

Cause of death: angular overlap as a result of a flawed experiment

```
Spacegroup name           P21
Unit cell parameters      75.185 110.553 173.212 90.000 90.233 90.000
Wavelength                1.00001 A
```

- Near orthorhombic cell, with a long c axis and the c* axis at 87.5 degrees to the Omega (rotation) axis
- 1-degree images (*on a Pilatus 6M in October 2016 … !*) causing reflection overlap for Omega ranges where the c* axis is close to parallel to the beam.

Check: simulate the diffraction pattern with 0.1-degree image width and count the number of predicted reflections pairs that are within +-10 pixels (0.172 mm) and images (0.1 deg) of each other: this predicts the problematic image rages exactly

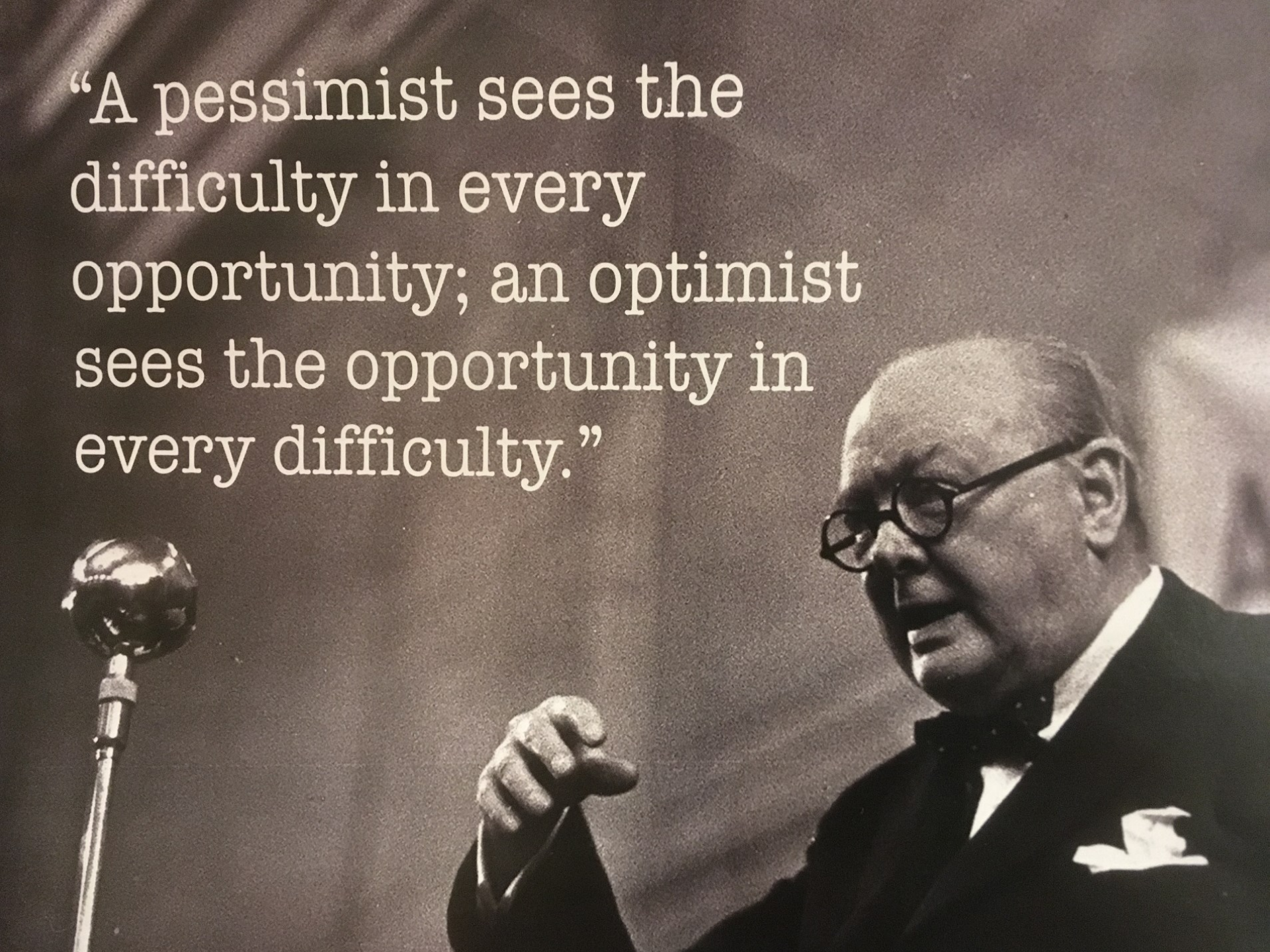


Outlook: we need not just to get **better at archiving** whatever data have been collected, but **better at collecting** those data in the first place!

GΦL
Global Phasing Limited

Conclusions

- The re-use of archived diffraction data plays a major role in our activities within the “virtuous circle” of iterative mutual improvement of software and revised final results made possible by the availability of data *upstream* of those deposited with PDB entries.
- While raw diffraction images are an indispensable source of such data, an enriched version of the currently archived unmerged data has the potential of enabling much of the achievable improvements in both software and final results, with the advantage that their availability in standardised form would circumvent the current rough edges of the direct re-use of raw data.
- A collective activity in this direction has been running under the auspices of the SubGroup on Data Collection and Processing of the PDBx/mmCIF Working Group to implement extensions to the mmCIF dictionary aimed at supporting the deposition and archiving of such enriched reflection data for both rotation and single-shot serial (SSX and SFX) crystallography.
- **It is imperative that this Commission should consider endorsing this effort and its goals, so as to speed up the approval of allocating internal PDB resources to actually carry out the tasks of harvesting and archiving the contents of such enriched files, once deposited.**


“A pessimist sees the difficulty in every opportunity; an optimist sees the opportunity in every difficulty.”

GΦL
Global Phasing Limited

Acknowledgements

- Global Phasing colleagues
- Peter Keller
- Rasmus Fogh
- Wlodek Paciorek
- Claus Flensburg
- Clemens Vonrhein
- Andrew Sharff
- Ian Tickle
- Marcin Wojdyr

- EMBL-Hamburg / PETRA III
- Gleb Bourenkov
- Ivars Karpics
- MPI Goettingen
- Ashwin Chari

- The MXCuBE and ISPyB Collaborations

- The PDBx/mmCIF Subgroup on Data Collection and Processing, especially
- John Westbrook (†)
- Ezra Peisach
- Stephen Burley
- Aaron Brewster
- David Waterman

- The users of our software who sent feedback and shared data

GΦL
Global Phasing Limited

Last but not least:
The Global Phasing Consortium