Requirements in Automatic Data Processing

C. Vonrhein, C. Flensburg & G. Bricogne
Global Phasing Ltd.

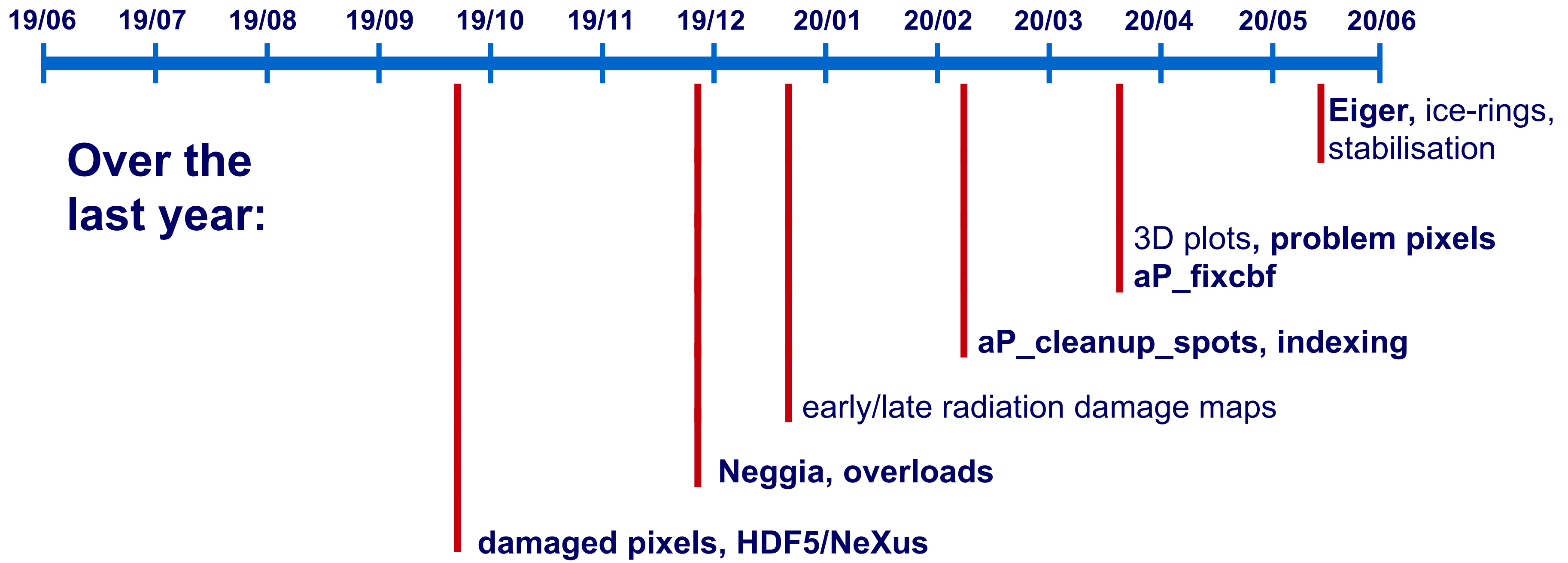
Workshop on MX raw image data formats, metadata and validation
22nd August 2020

GΦL
Global Phasing Limited

Release timeline of autoPROC software

https://www.globalphasing.com/autoproc/news.html

Over the last year:



Some topics of different autoPROC releases: pixel problems, detector/image formats, indexing, reporting, automation (of course) etc.

GΦL
Global Phasing Limited

HDRMX

- Over the last couple of years (decades?), at every user/developer meeting and workshop with data processing as a topic: **"Let's talk/moan about image headers and formats ..."**

- d*TREK, fullCBF, imgCIF
- ADSC, marCCD, Bruker ...
- mini-cbf
- **HDF5 and NeXus (NXmx) format**
- **Eiger detectors**

A1. Simplified NXmx layout

A tree representation of this simplified NXmx layout is available at

http://hdrmx.medsbio.org/gold2/NXmx_Gold_Standard.jpg

```
<?xml version="1.0" encoding="UTF-8"?>
<definition name="NXmx" extends="NXobject" type="group"
    category="application"
    xmlns="http://definition.nexusformat.org/nxdl/3.1"
    xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
    xsi:schemaLocation="http://definition.nexusformat.org/nxdl/3.1 ../nxdl.xsd"
    >
        <group:NXentry>
            <field:title type="NX_CHAR" optional />
            <field:start_time type="NX_DATE_TIME" />
            <field:end_time type="NX_DATE_TIME" optional />
            <field:end_time_estimated type="NX_DATE_TIME" />
            <field:definition />

            <group:NXdata>
                <field:data type="NX_NUMBER" recommended />
            </group:NXdata>

            <group:NXsample>
                <field:name type="NX_CHAR" />
                <field:depends_on type="NX_CHAR" />
                <group:"NXtransformations" recommended />
                <field:"temperature" units="NX_TEMPERATURE" optional />
            </group:NXsample>
```

Let's talk about "image data content" ... much more interesting (?)

Global Phasing Limited

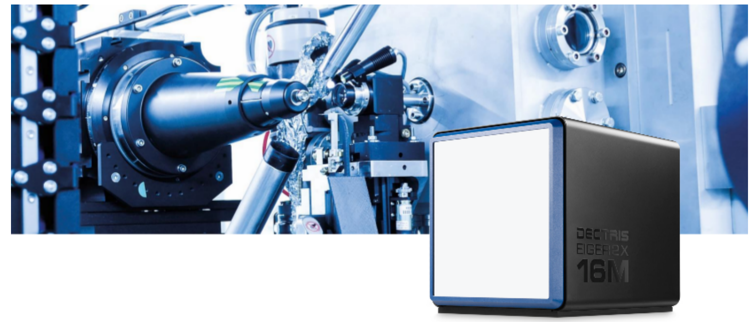
Eiger(2) data: 16-bit vs 32-bit

3.1. Specifications

3.1.2. Detector

Table 3.2: Technical Specifications

Image bit depth	16 bit or 32 bit
Readout bit depth	12 bit



Technical Specifications

EIGER2 X 16M

3.1. Specifications

3.1.1. Detector

Table 3.1: Technical Specifications

Image bit depth	32 bit
Readout bit depth	16 bit

filewriter (HDF5) or stream interface

uint16 or uint32	uint32
0..65535 0..4294967295	0..4294967295

applications mostly want int32

-2147483648..2147483647

Global Phasing Limited

Eiger data: 16-bit vs 32-bit

- Eiger/Eiger2 data typically provided to applications via one of
  - Dectris/Neggia **plugin**
  - Diamond/Durin **plugin**
  - **Conversion** from HDF5 to mini-cbf:
    - H5ToXds (Dectris)
    - https://github.com/biochem-fan/eiger2cbf
    - hdf2mini-cbf (GPhL)
  - writing from **Stream interface** into mini-cbf (application X?)

- Each application needs to handle
  - uint16 vs uint32 data
  - UINT16_MAX and UINT32_MAX markers:
    - unsigned int data doesn't provide negative markers
    - 0 is a valid pixel count
    - Eiger/Eiger2 firmware uses UINT{16,32}_MAX marker
  - conversion to int32 (signed 32-bit integer)
- **before actual processing data**

Caveat: this is our understanding and might not be a complete picture ... but we had lots of conversations with experts (Dectris, APS, Petra-III, DLS, XDS etc) that seem to confirm the gist this.

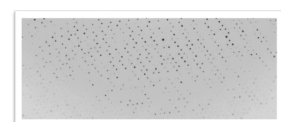
Eiger data: 16-bit vs 32-bit

- **observation**: default autoPROC run on HDF5 data (using the Dectris/neggia plugin) gives a large number of very strong spots in COLSPOT
- **convert** HDF5 to mini-cbf
- **looking at** these **images** (using e.g. GPX2 or ADXV): there are no strong spots at all it seems. **Odd**!
- since raw data is stored (compressed) inside HDF5 container we need a reliable tool to convert this into a format we can visualise/inspect
- we provide
  - **hdf2mini-cbf**:
    - should support all kind of HDF5 files/formats out there
    - but needs constant checks and updates (because HDF5 files/formats can change regularly)
  - **cbf2ijk**: write ASCII version of x,y,value

DECTRIS

PILATUS CBF

Header Specification



Version 1.4

Global Phasing Limited

Eiger data: 16-bit vs 32-bit

Dectris/Neggia (autoPROC)

```
NUMBER OF STRONG PIXELS EXTRACTED FROM IMAGES   1509967
NUMBER OF DIFFRACTION SPOTS LOCATED               66279
IGNORED BECAUSE OF SPOT CLOSE TO UNTRUSTED REGION  3431
WEAK SPOTS OMITTED                                23352
NUMBER OF DIFFRACTION SPOTS ACCEPTED              39496
```

Dectris/Neggia (XDSme)

```
NUMBER OF STRONG PIXELS EXTRACTED FROM IMAGES   1509967
NUMBER OF DIFFRACTION SPOTS LOCATED               66279
IGNORED BECAUSE OF SPOT CLOSE TO UNTRUSTED REGION  3431
WEAK SPOTS OMITTED                                23352
NUMBER OF DIFFRACTION SPOTS ACCEPTED              39496
```

DLS/Durin (autoPROC)

```
NUMBER OF STRONG PIXELS EXTRACTED FROM IMAGES   1540328
NUMBER OF DIFFRACTION SPOTS LOCATED               67110
IGNORED BECAUSE OF SPOT CLOSE TO UNTRUSTED REGION  3510
WEAK SPOTS OMITTED                                10068
NUMBER OF DIFFRACTION SPOTS ACCEPTED              53532
```

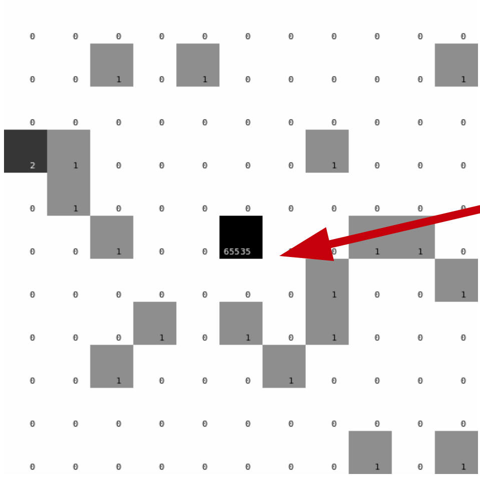
hdf2mini-cbf

```
NUMBER OF STRONG PIXELS EXTRACTED FROM IMAGES   1540328
NUMBER OF DIFFRACTION SPOTS LOCATED               67110
IGNORED BECAUSE OF SPOT CLOSE TO UNTRUSTED REGION  3510
WEAK SPOTS OMITTED                                10068
NUMBER OF DIFFRACTION SPOTS ACCEPTED              53532
```

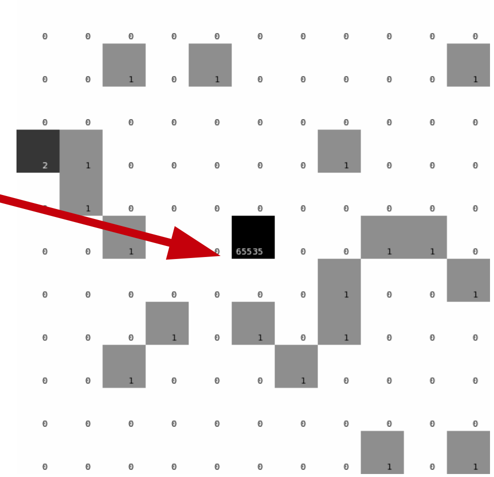
Same (16-bit) dataset

HDF5 → mini-cbf: know your converter

H5ToXds



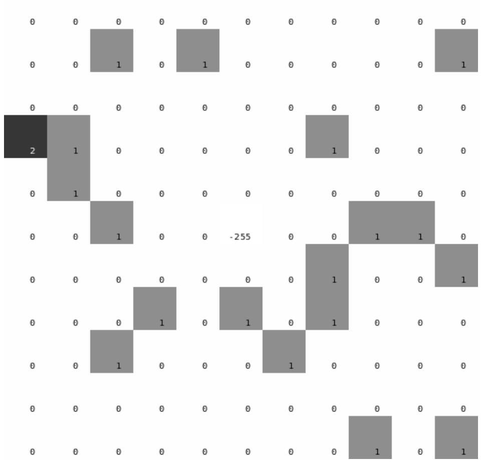
eiger2cbf



incorrect

hdf2mini-cbf

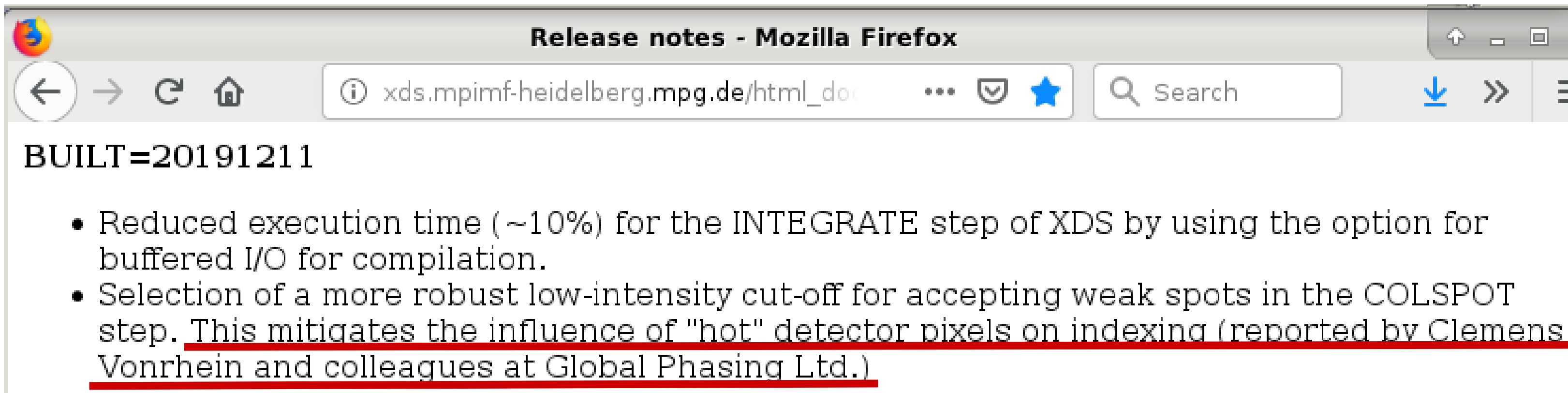
-255



- there *should* be no need for converters: always best to process data as close to original as possible (but *extractors* still needed)
- HDF5 archives sometimes cumbersome
- mini-cbf nice and simple: allows comparison between Albula, ADXV, GPX2, various libraries and tools

GΦL
Global Phasing Limited

Feedback to XDS developers



BUILT=20191211

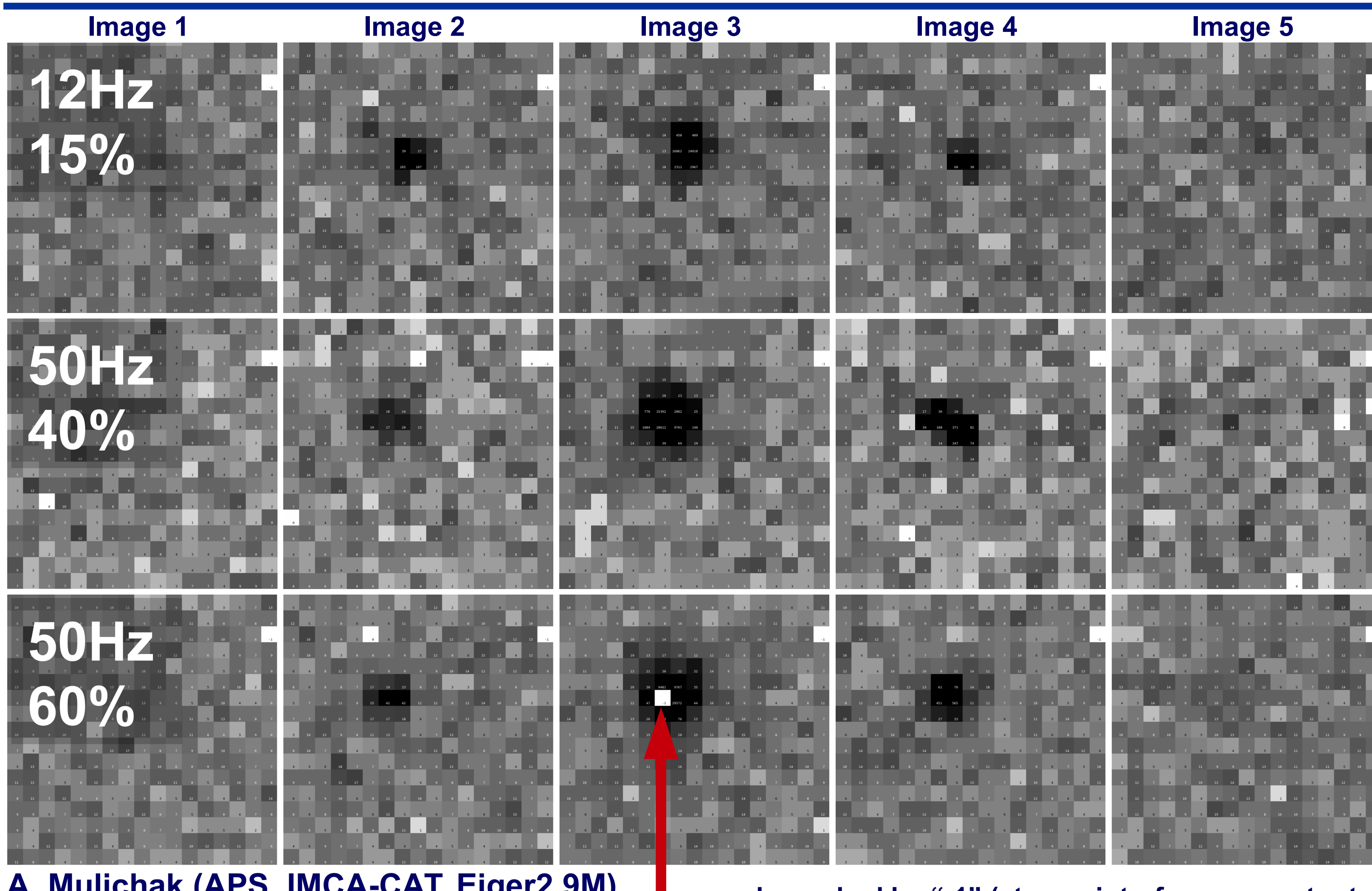
- Reduced execution time (~10%) for the INTEGRATE step of XDS by using the option for buffered I/O for compilation.
- Selection of a more robust low-intensity cut-off for accepting weak spots in the COLSPOT step. This mitigates the influence of "hot" detector pixels on indexing (reported by Clemens Vonrhein and colleagues at Global Phasing Ltd.)

- Moving from cut-off based on mean (can be influenced by outliers, i.e. hot and too strong pixels) to median-based criteria
- Obviously still better to exclude hot/damaged/flickering pixels from entering (XDS) computations in the first place.

- Exclusion/marking of pixels needs to be done correctly: **what “marker” to use?**

GΦL Global Phasing Limited

UINT32_MAX: hidden “treasures”



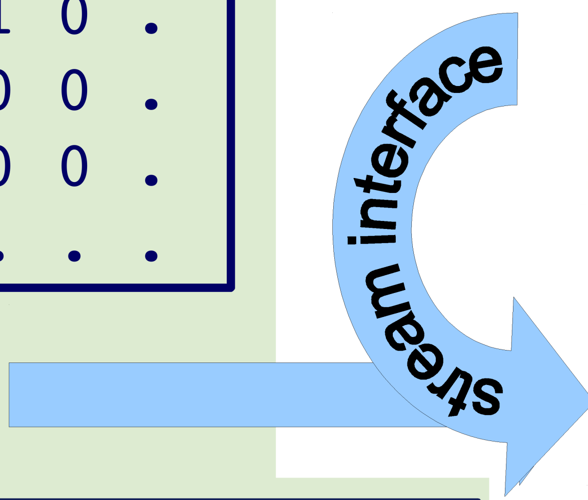
A. Mulichak (APS, IMCA-CAT, Eiger2 9M)

wrongly marked by “-1” (stream interface converter to mini-cbf)

Eiger: stream to mini-cbf conversion

pixel mask

```
.  .  .  .  .  .
.  0  0  1  0  .
.  0  2  0  0  .
.  4  0  0  0  .
.  .  .  .  .  .
```




```
bit 0 ( 1): gap (pixel with no sensor)
bit 1 ( 2): dead
bit 2 ( 4): under responding
bit 3 ( 8): over responding
bit 4 (16): noisy
bit 5 (32): -undefined-
bit 6 (64): pixel is part of a cluster of problematic pixels
            (bit set in addition to others)
bit 7(128): -undefined-
bit 8(256): user defined mask (e.g. around beamstop)
bits 9-30: -undefined-
bit 31: virtual pixel (corner pixel with interpolated value)
```

UINT32_MAX:

pixel_mask >0:
-pixel_mask

pixel_mask =0:
Count_cutoff+1

XDS: overloaded vs dead pixels

XDS Input Parameters - Mozilla Firefox

xds.mpimf-heidelberg.mp

MINPK=

Defines the minimum required percentage of observed reflection intensity. The missing intensity is estimated from the learned profiles. If less than MINPK % is observed, the reflection will be discarded.

Example: MINPK=75.0

The default value of 75% works fine and hardly needs to be changed.

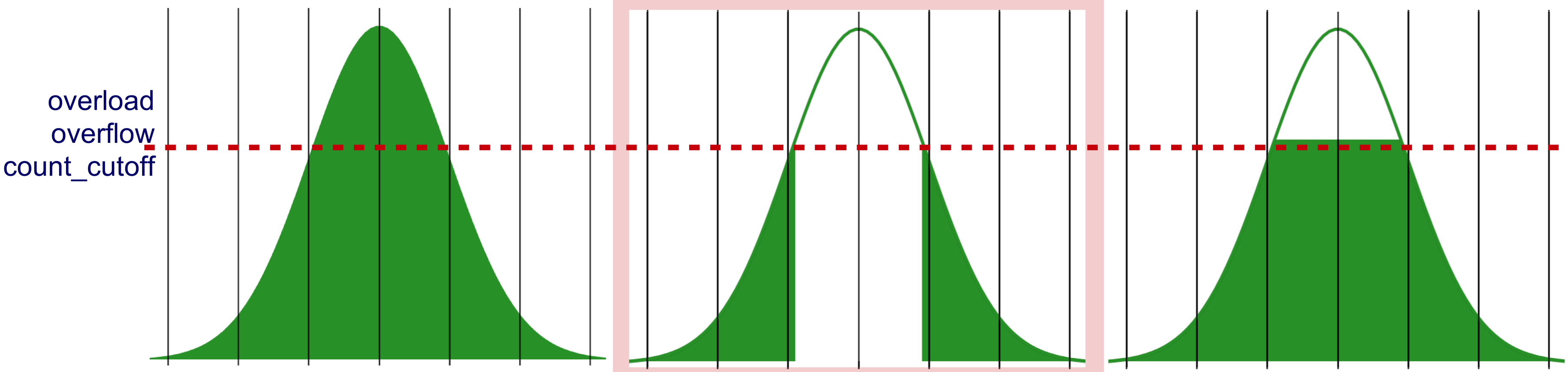
XDS Input Parameters - Mozilla Firefox

xds.mpimf-heidelberg.mp

OVERLOAD=

If a pixel contents exceeds this maximum value the pixel is overloaded; a reflection is overloaded if it includes one or more overloaded pixels. In the "INTEGRATE" step overloaded reflections are excluded from the determination of reference profiles. Otherwise they are treated like all other reflections and saved on the output file INTEGRATE.HKL. In the "CORRECT" step of XDS overloaded reflections are excluded from the final output because their integrated intensities are incorrect.

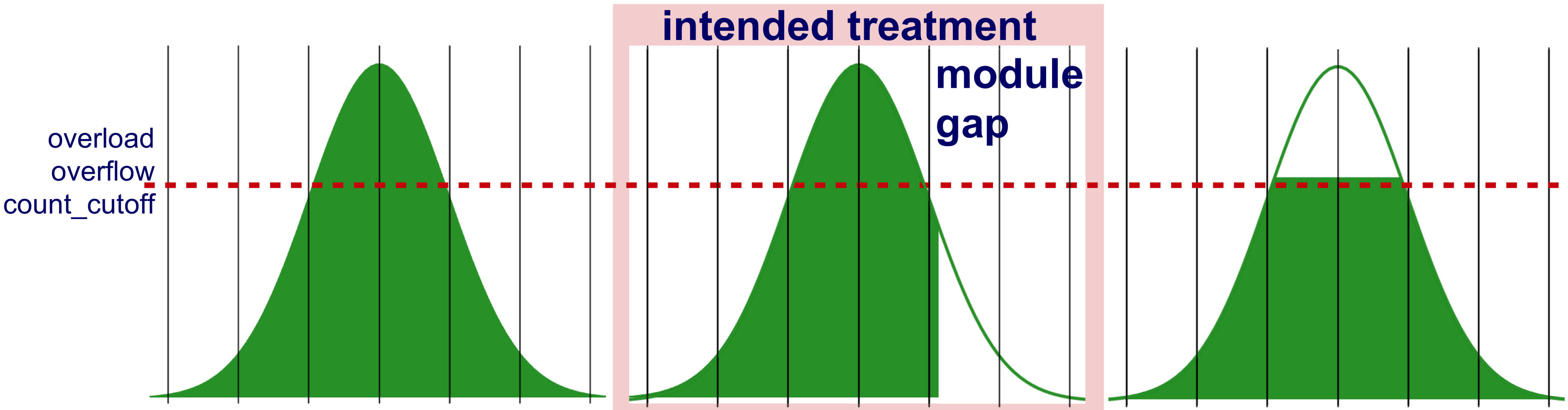
XDS: differences depending on marking



COLSPOT (spot search)	**used**	**used**
INTEGRATE (integration)	**profile formation intensity estimated**	**~~profile formation~~ profile fitted**
CORRECT (scaling/correction)	**used (estimated intensity)**	**excluded (untrustworthy)**

---------------- strongest reflections ----------------

XDS: differences



	intended treatment	
COLSPOT (spot search)	**used**	**used**
INTEGRATE (integration)	**profile formation intensity estimated**	**~~profile formation~~ profile fitted**
CORRECT (scaling/correction)	**used (estimated intensity)**	**excluded (untrustworthy)**
	random reflections	**strongest reflections**

GΦL
Global Phasing Limited

aP_fixcbf: for Eiger(2) data as mini-cbf

```
5V01 WARNING : a total of 77274 pixels have been reset in 515
5V01           out of 1440 images. This corresponds to an average
5V01           of 54 pixels for all images (or 150 pixels in all
5V01           affected images).

6MOL WARNING : a total of 13644 pixels have been reset in 303
6MOL           out of 1800 images. This corresponds to an average
6MOL           of 7.6 pixels for all images (or 45 pixels in all
6MOL           affected images).

6NRQ WARNING : a total of 7835 pixels have been reset in 151 out
6NRQ           of 925 images. This corresponds to an average of
6NRQ           8.5 pixels for all images (or 52 pixels in all
6NRQ           affected images).

6UKG WARNING : a total of 4724 pixels have been reset in 1599
6UKG           out of 1600 images. This corresponds to an average
6UKG           of 3.0 pixels for all images (or 3.0 pixels in all
6UKG           affected images).

6DHW WARNING : a total of 4566 pixels have been reset in 94 out
6DHW           of 360 images. This corresponds to an average of
6DHW           13 pixels for all images (or 49 pixels in all
6DHW           affected images).

6P7P WARNING : a total of 290 pixels have been reset in 136 out
6P7P           of 900 images. This corresponds to an average of
6P7P           0.322222 pixels for all images (or 2.1 pixels in
6P7P           all affected images).
```

APS, 21-ID-D

APS, 23-ID-B

APS, 23-ID-B

APS, 22-ID

APS, 21-ID-D

APS, 24-ID-E

Global Phasing Limited

6UKG (~3 pixels/image reset)

data as-is

	Overall	InnerShell	OuterShell
Low resolution limit	57.989	57.989	1.144
High resolution limit	1.083	3.139	1.083
Rmerge (all I+ & I-)	0.082	0.073	0.828
Rmeas (all I+ & I-)	0.088	0.079	0.943
Rpim (all I+ & I-)	0.033	0.029	0.444
Total number of observations	748155	37463	23422
Total number unique	107912	5395	5397
Mean(I)/sd(I)	12.0	23.6	1.6
Completeness (spherical)	84.2	99.0	27.7
Completeness (ellipsoidal)	88.9	99.0	36.9
Multiplicity	6.9	6.9	4.3
CC(1/2)	0.996	0.992	0.611

after aP_fixcbf

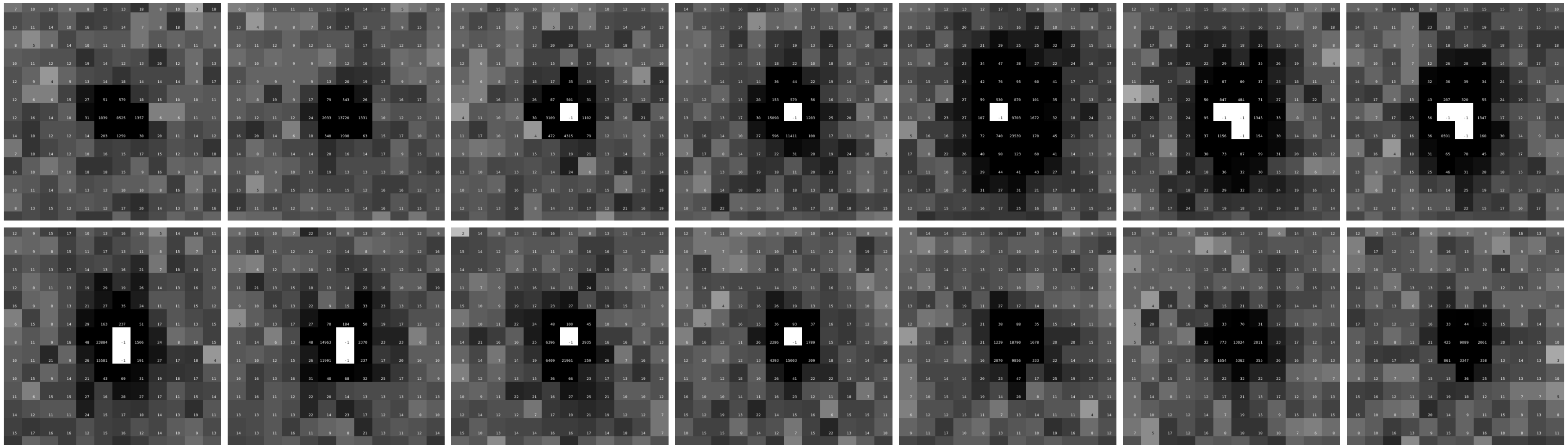
	Overall	InnerShell	OuterShell
Low resolution limit	57.989	57.989	1.144
High resolution limit	1.083	3.133	1.083
Rmerge (all I+ & I-)	0.080	0.070	0.828
Rmeas (all I+ & I-)	0.087	0.076	0.943
Rpim (all I+ & I-)	0.032	0.028	0.444
Total number of observations	747774	37398	23398
Total number unique	107884	5393	5395
Mean(I)/sd(I)	12.0	23.8	1.6
Completeness (spherical)	84.1	98.4	27.7
Completeness (ellipsoidal)	88.9	98.4	36.9
Multiplicity	6.9	6.9	4.3
CC(1/2)	0.996	0.994	0.610

Only small differences in overal scaling/merging statistics after fixing diffraction data.

But ...

GΦL
Global Phasing Limited

UINT32_MAX: hidden "treasures"



G. Bourenkov (Petra-III)

- **aP_fixcbf** is more an **analysis tool** and not intended as an integral part of a "production" pipeline
- there will be marginal differences in scaling/merging statistics

- main point: **overloaded pixels/reflections become invisible (important information to decide on potential adjustment of collection strategy, detector settings, speed, flux etc.)**

WARNING

1191 overloaded reflections out of 769541 total - which is rather unexpected for an EIGER detector (we would expect data collected with low dose and high multiplicity). There could be good reasons for this, but you might want to check (e.g. with the beamline staff) if (1) the pixel mask is set/applied correctly, (2) some damaged ("hot", "flickering" or "dead") pixels are not yet masked, (3) the transmission was larger than necessary or (4) the data collection speed was too fast. There could be other reasons as well ... the main point is that this doesn't look good and needs investigating.

6UKG: WARNING from autoPROC (summary.html)

GΦL
Global Phasing Limited

Summary, status and outlook

- Will we still be talking about meta-data formats in 2021?
- Hopefully we'll be talking "only" about pixel data content by then.
- Interactions with beamlines and external developers important (e.g. via HDRMX, MXCuBE, ISPyB)!
- Interaction with power users crucial for feedback, real-life testing and planning (EMBL/CRIMS, GPhL consortium members etc)!

- More automation (autoPROC) expected especially in decision making about poor image ranges.

Thanks to a lot of people: lots of Consortium members, beamline and synchrotron staff, developers & GPhL colleagues

H. Bernstein, T. Bertrand, G. Bey, M. Blaesse, G. Bourenkov, G. Bunkoczi, R. Byrne, I. Cornaciu, J. Dias, K. Diederichs, P. Evans, G. Fischer, T. Fischmann, M. Haffke, S. Harris, W. Kabsch, J. Key, E. Krissinel, J. Kopec, I. Korndoerfer, M. Kroemer, A. Kuglstatter, A. Lammens, P. Legrand, M. Lehmann, S. Liu, D. Logan, K. Longenecker, J. Marquez, M. Mathieu, P. McEwan, L. Miallau, A. Mulichak, J. Murray, D. Musil, R. Nolte, M. Rappas, J. Read, D. Reinert, P. Rowland, P. Rucktooa, M. Rudolph, J. Sack, M. Savko, C. Schleberger, H. Schreuder, S. Sheriff, O. Svensson, M. Swan, E. TerHaar, J. Thorpe, Y. Wang, D. Waterman, T. Weinert, M. Weiss, G. Winter, J. Wojdyla, D. Zeyer ... many more