Project information

Project full title	LEAPS pilot to foster open innovation for accelerator-based light sources in Europe
Project acronym	LEAPS-INNOV
Grant agreement no.	101004728
Instrument	Research and Innovation Action (RIA)
Duration	01/04/2021 – 31/03/2025
Website	

Deliverable information

Deliverable no.	D7.4
Deliverable title	Best practices and guidelines for applying data reduction and compression at LEAPS facilities
Deliverable responsible	HZDR
Related Work-Package/Task	WP7, Task 7.4
Type (e.g. Report; other)	Report
Author(s)	Peter Steinbach, David Pennicard, Zdenek Matej, Vincent Favre-Nicolin, Nicolas Soler, Markus Janousch, Krzysztof Madura, Francesc Alted
Dissemination level	
Document Version	
Date	28.05.2024
Download page	

Document information

Version no.	Date	Author(s)	Comment
1.0	28.05.2024	P. Steinbach, D. Pennicard, Z. Matej, V. Favre-Nicolin, N. Soler, M. Janousch, K. Madura, F. Alted	

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Best practices and guidelines for applying data reduction and compression at LEAPS facilities

Introduction

Compression of recorded data is a crucial aspect of photon sciences, particularly in the fields of imaging and diffraction. In these disciplines, enormous volumes of data are generated through advanced techniques like tomography, multispectral imaging, scanning diffraction imaging and serial crystallography. Effective data compression is essential for several key reasons:

1. Storage Space and Recording rate: Raw data from photon science experiments can quickly consume vast amounts of storage space. In addition it is challenging and resource demanding to record raw data at the rate they are produced by detectors nowadays. Compression reduces the size of the data, making it more manageable and cost-effective to store and archive.

2. Data Transfer: Compressed data is easier and faster to transfer over networks, facilitating collaboration between researchers and enabling efficient data sharing.

3. Processing Speed: Smaller file sizes resulting from compression allow for faster data processing, analysis, and visualization, accelerating the pace of research and discovery.

4. Signal Extraction: Compression algorithms can help extract meaningful signals from noisy data, enhancing the quality and interpretability of the results.

5. Long-Term Preservation: Compressed data formats ensure the longevity of valuable scientific data, making it accessible and usable for future research endeavors.

Data compression plays a pivotal role in photon sciences by reducing data volume, expediting data transfer and processing, improving signal quality, and enabling long-term data preservation and accessibility.

In this document, we want to summarize our findings in LEAPS-INNOV WP7 to tackle these challenges and opportunities outlined above in the photon sciences. Most of all, this document is structured as follows: we first report on performance of new compression schemes, then we discuss how these schemes can be integrated into experimental setups; we then also consider how to achieve user community; we finalize the report by providing concrete practices for present day applications.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Evaluating the performance of new compression schemes

For evaluating compression methods, we require a quantitative framework for evaluating their performance, speed, and quality of compression. The choice of compression algorithm depends on the specific requirements of the application, balancing factors such as compression ratio, speed, data fidelity, and hardware efficiency. ***Compression Ratio*** is a fundamental metric that quantifies the effectiveness of a compression method. It is calculated as the ratio of the original file size to the compressed file size. ***Compression speed*** refers to the time taken to compress the original data into a compressed format. It is measured in megabytes or gigabytes per second (MB/s or GB/s). Faster compression speeds are desirable, especially when dealing with large datasets or real-time applications. ***Decompression speed*** is the time taken to decompress the compressed data back to its original format. Like compression speed, it is also measured in MB/s or GB/s. Faster decompression speeds are important for quick data access and processing. Another key distinction is whether the compression method is *lossless* or *lossy*. ***Lossless compression*** ensures that the decompressed data is identical to the original, while ***lossy compression*** allows for some data loss to achieve higher compression ratios. The choice depends on the application's tolerance for data accuracy.

Collected Imaging Datasets for this study

Based on the challenges outlined above, our work culminated on the following datasets. We focused on imaging techniques such as tomography because imaging data is currently typically uncompressed, and occupies a disproportionately large amount of disk space at many of our facilities. For each of these datasets, we were able to establish contact with personnel who were involved in creating these datasets.

The ROFEX dataset

The high-performance ROFEX (**RO**ssendorf **F**ast **E**lectron beam **X**-ray tomography) imaging technique has been developed at Helmholtz-Zentrum Dresden-Rossendorf for the noninvasive investigation of dynamic processes. Purpose-built for the analysis of multiphase flows in elongated test sections, various research projects out of this field could already benefit from the ROFEX technology. Beyond that it can be applied to diverse other applications such as for instance nondestructive testing.

The dataset depicted in figure 1 left was provided as a hdf5 file as produced by Matlab. Each file consisted of 15000 frames of observations. The dataset itself used in this analysis was encoded as 32 floating point grayscale. The data comprised 15000 frames of WxH = 256x256 pixels each. The dataset of figure 1 right consists of 1024 frames of shape WxH = 256x256 pixels each. The grayscale pixel intensities are encoded as 16 bit unsigned integer values.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

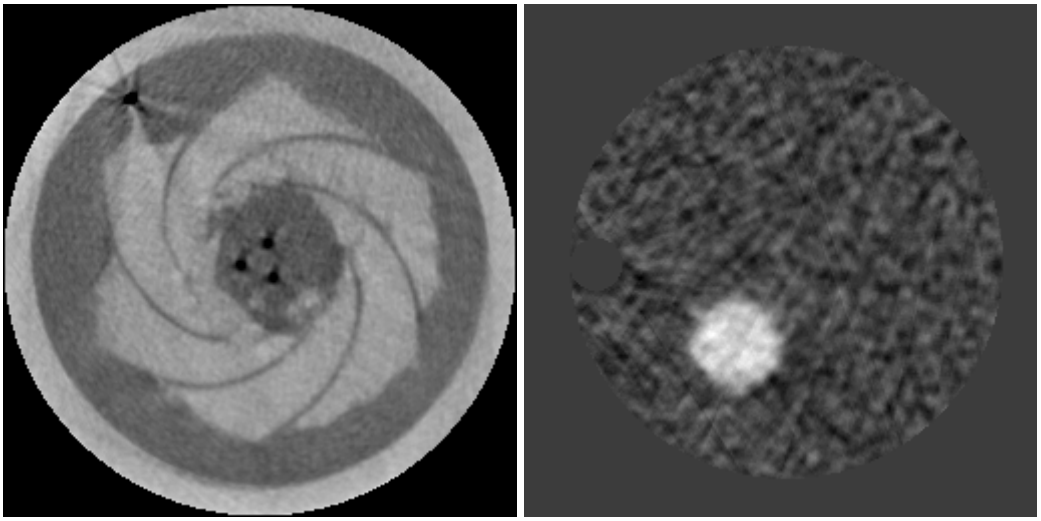

Figure 1: Example frames of a ROFEX multi-phase flow experiment (left) and a single bubble experiment (right). Both images were extracted from a dataset after reconstruction and postprocessing.

The PSI evolving magma dataset

In the associated article to these datasets (MI04_02, Pistone 2021), the authors investigate the effect of lossy compression of original X-ray projections onto the final tomographic reconstructions. The dataset in use was recorded at the TOMCAT experiment at Paul Scherrer Institut. The beamline for TOmographic Microscopy and Coherent rAdiology experimentTs (TOMCAT) is operated by the X-ray Tomography Group and offers cutting-edge technology and scientific expertise for exploiting the distinctive features of synchrotron radiation for fast, non-destructive, high resolution, quantitative investigations on a large variety of samples.

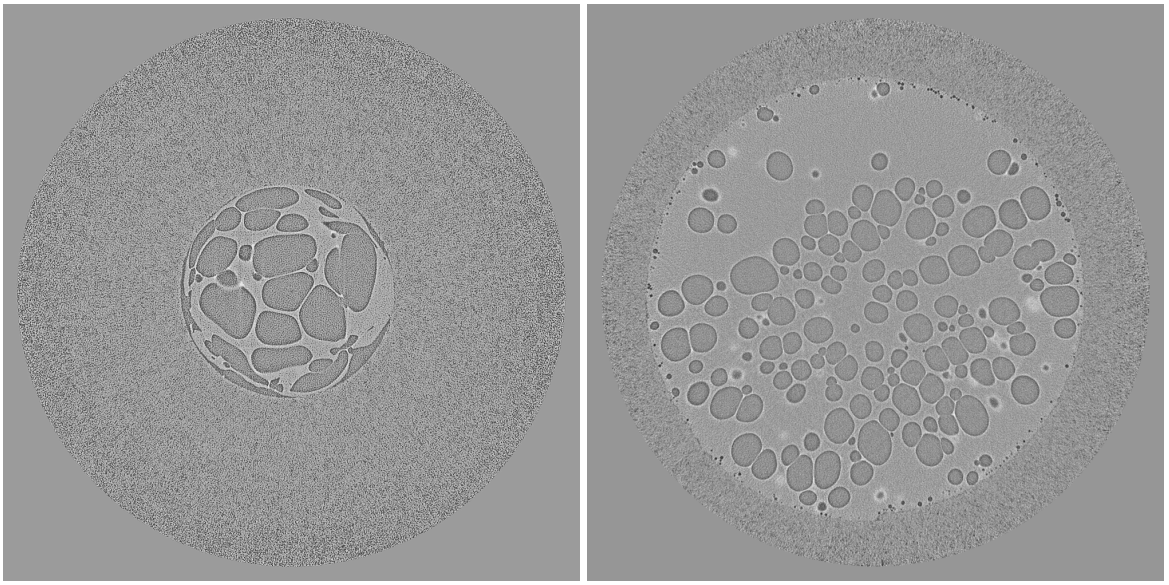

Figure 2: Example frames of a PSI evolving magma dataset. Both images were extracted from a dataset after reconstruction.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

The raw dataset contained 751 frames encoded as 16 bit unsigned int. Each frame had a shape of WxH = 1900x1008 pixels. The full volume of this dataset comprises 2.743 GB and represents one measurement by the TOMCAT detector. This dataset and many others are available at https://doi.psi.ch/detail/10.16907/05a50450-767f-421d-9832-342b57c201af

Lossless Compression

The project aimed to explore lossless data compression techniques for tomography data, particularly the PSI magma dataset. We encountered challenges in finding additional complementary datasets due to data portal unavailability like data.panosc.org and data owner hesitancy regarding proprietary concerns. The PSI magma dataset served as a crucial starting point for our work.

We applied existing lossless compression techniques from blosc2 to evaluate their effectiveness. However, we faced difficulties in reproducing the full tomography reconstruction pipeline due to a lack of documentation of employed parameters used in standard software packages. This highlighted a broader reproducibility issue in tomography data handling. We henceforth emphasized to our communities the importance of comprehensive documentation of reconstruction parameters to improve reproducibility and facilitate downstream innovation in data compression.

Modern compression orchestrators like blosc2 offer a wide range of tunable parameters to tailor compression behavior to specific data types. Finding the optimal parameter set is crucial for achieving maximum compression ratios. By compressing individual sinogram layers with blosc2's BloscLZ algorithm, we achieved a compression ratio of 2.12x, with ingest bandwidths of up to 1 GB/s. This justified an existing heuristic, that lossless compressions schemes can attain compression ratios around two. In practice and given modern imaging sensor technology, this is however not enough.

Through collaboration with the blosc2 development team, we utilized ironArray's Btune Studio[1] parameter optimization tool, which employs deep learning techniques to find optimal parameter sets specific for datasets following similar data patterns. This analysis surfaced parameter sets which can attain higher compression ratios of up to 3.98x and ingestion bandwidths of 9.23 GB/s in a fully lossless compression configuration. Trading off some compression ratio further increased ingestion bandwidths to 23.24 GB/s for individual sinograms of the PSI magma dataset. Both of these findings came much as a surprise for everyone involved.

Given the availability of high bandwidth pixel-array detectors like Jungfrau (Leonarski and Brückner 2022), experiments will soon be faced with data production bandwidths of 17 GB/s until up to 30 GB/s. Our findings demonstrate the potential of lossless compression in reducing data volumes by up to 4x, which is especially relevant with the emergence of these high-bandwidth pixel-array detectors. The automated and AI-guided search through codec configuration parameters highlights the

[1] https://ironarray.io/btune

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

importance of comprehensive analysis and benchmarking for production environments in the photon sciences.

Overall, our experience underscores the challenges in data sourcing and reproducibility and the need for improved data sharing practices to facilitate future innovations in data compression and storage optimization. FAIR principles should be considered a required prerequisite for sharing photon science datasets, but reproducibility of downstream analysis pipelines must be considered of equal importance.

Lossy Compression for tomography

We explored lossy compression techniques as an alternative to lossless compression, which has limited applicability in real-world scenarios. Lossy compression reduces file size by removing less noticeable details, similar to reducing image quality. While lossless compression has its benefits, we sought higher compression ratios and similar compression bandwidths to match the demands of automated beamline experiments operating continuously. We identified the stochasticity of pixel intensities (due to sensor noise for example) in image data as one limiting factor for compression yield.

We applied denoising algorithms to reduce this noise and improve compression. Our study used Butterworth and bm3d denoising on the PSI magma dataset sinograms, resulting in improved compression ratios of up to 2.4x. We also analyzed the reconstructed tomograms and found that compression codecs struggled to achieve ratios above 1.5x due to intensity distortions stemming from the tomogram reconstruction. However, denoising algorithms mitigated this issue, achieving a maximum compression ratio of 2.4x. While this proves the point that noise removal and image restoration can support image compression, the achieved compression ratio proves far remote from what is required by modern sensors and storage infrastructure.

For this reason, we implemented a uniform quantization of the ROFEX dataset to reduce the 16-bit intensity range of recorded imagery to 8-bit integer encoding, resulting in a 2x reduction in file size on top of what is achieved by compression approaches mentioned earlier. We further enhanced this approach for a single-bubble dataset by thresholding with Otsu's method. This thresholding gives rise to a more dynamic encoding scheme, where anything considered background (low intensities) is encoded in 2 bits while the remaining (signal) is encoded with 6 bits. This simple and fast approach, abbreviated by TQC, achieved a remarkable compression ratio of 18.34x.

We addressed the challenge of evaluating the impact of lossy compression on scientific data. Traditional image quality metrics like PSNR or SSIM are used to assess the effect of lossy image compression codecs on scientific imagery. But this may not provide meaningful insights for scientists and their subsequent uses of the data. By collaborating with dataset owners, we obtained access to downstream analyses and software, allowing us to compare the original data with decompressed

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

data. We computed the Normalized Root Mean Square Error (NRMSE) to quantify the differences in 15 physics variables per pixel which are used by scientists to perform scientific inferences with the data obtained. Our TQC lossy compression scheme resulted in NRMSE values below 1% in the permille regime, demonstrating minimal impact on the data's scientific value.

Our work highlights the potential of lossy compression in achieving higher compression ratios while maintaining data integrity for scientific use. By providing concrete evaluations of compression techniques' effects on downstream analyses, we aim to make lossy compression more approachable and understandable for scientists. This again underpins the importance of reproducible downstream analysis pipelines honoring the FAIR principles. Moreover, our enterprise demonstrates that an in-depth understanding of the data collected is essential to foster high compression ratios at the required modern data production bandwidths.

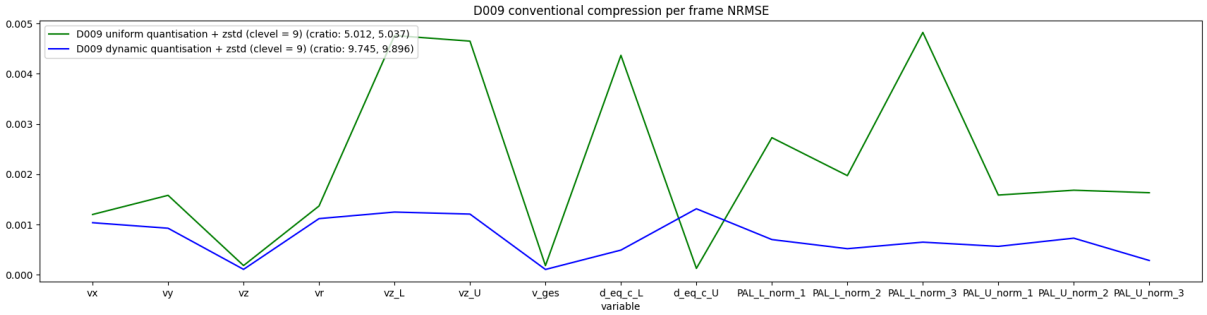

Figure 7: Analysis of the effect of lossy compression on the single bubble ROFEX dataset. The figures compare the NRMSE difference between the original data versus a decompressed version of it. We compare naive quantisation (blue) and the TQC algorithm (green). Zstandard was used as the eventual data encoder.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Integrating new compression schemes into experiments

As pointed out earlier, while our facilities face the data deluge, lossy compression becomes a mandatory step after data collection for some techniques like micro-Computed tomography (µ-CT), where only little benefit is gained from lossless compression. However, in order to be achieved, lossy compression must be applied respecting the following conditions:

1. The loss in data quality upon compression must be small enough so as to result in minimal or unchanged scientific data quality, so that subsequent analysis steps (tomographic volume reconstruction, image segmentation, labeling, quantification etc..) remains unperturbed, as measured by the figures-of-merit discussed in the 'lossy compression' paragraph above.
2. Interoperability: compression and decompression must be compatible with current file formats and standards used in photon and neutron facilities, and ultimately be transparent to the scientific users.

Our collaboration with the BLOSC team has resulted in solutions addressing these two points: JPEG2000 was chosen as the algorithm of choice for lossy compression of tomographic data due to its acceptance by the tomographic community[2]. After evaluating several codecs implementing jpeg 2000, the BLOSC team integrated the open source libraries OpenJ2HK[3] and Grok[4] into the Blosc2 meta compressor. Grok has been retained as the recommended option due its support for 16-bit gray images.

The solution proposed to the µ-CT community using Python is therefore to perform lossy compression of projection and reconstructed volume data with Blosc2/Grok and use hdf5plugin[5] to write the compressed datasets into the HDF5 format. These compressed datasets can be transparently read using h5py[6] provided that the hdf5plugin and blosc2-grok[7] Python packages are installed. In addition, this workflow can benefit from other features available with Blosc2 such as **direct-chunk writing** (faster access than using the regular HDF5 write process) and automatic **optimized reading** of n-dimensional slices (package b2h5py[8], allowing faster reading times). Examples and documentation on how to use this compression scheme have been provided by the BLOSC team on the web[9] [10]. Also, the ESRF team provided a working example involving h5py, b2h5py, hdf5plugin, Blosc2 and the grok codec all collaborating in creating and accessing and HDF5 file with JPEG2000 compressed images[11]. It should also be noted that the information regarding the codec used are embedded in the hdf5 files (at least, enough information to transparently uncompress the

[2] Marone, F., Vogel, J. & Stampanoni, M. (2020). J. Synchrotron Rad. 27, 1326-1338.
[3] https://github.com/osamu620/OpenHTJ2K
[4] https://github.com/GrokImageCompression/grok
[5] https://www.silx.org/doc/hdf5plugin/latest/
[6] https://www.h5py.org/
[7] https://pypi.org/project/blosc2-grok/
[8] https://github.com/Blosc/b2h5py
[9] https://www.blosc.org/posts/blosc2-lossy-compression/
[10] https://www.blosc.org/posts/blosc2-grok-release/
[11] https://gist.github.com/t20100/80960ec46abd3a863e85876c013834bb

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

data provided the plugins are installed), and more metadata could also be added as a NeXuS field or attribute to detail what exact options (and optimisation) where passed to the compression codec.

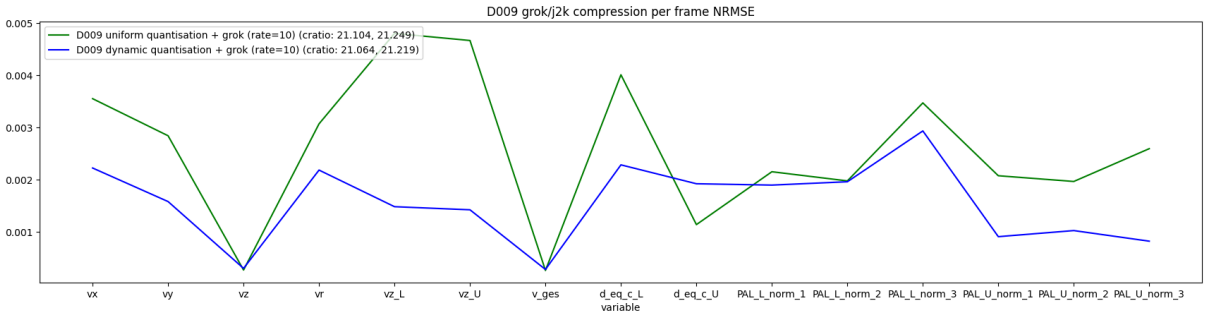

Figure 8: Analysis of the effect of lossy compression on the single bubble ROFEX dataset. The figures compare the NRMSE difference between the original data versus a decompressed version of it. We compare naive quantisation (blue) and the TQC algorithm (green). JPEG2000 (using the grok library) was used as the eventual data encoder using the blosc2 plugin elaborated in this project.

Figure 8 picks up the same pipeline as employed for figure 7 and makes the reconstructed tomograms subject to the grok/j2k encoder available through the blosc2-grok plugin. We can see how the physics reconstruction (quality described by the NRMSE metric on the y axis) remained at a comparable scale of up to 0.005, i.e. 0.5 %. Due to the advanced encoder features available using grok, the compression ratio (see legend of figure 8) jumped from a factor of 9.8 (in figure 7) to 21.2x. With this, we demonstrate the virtue of maintaining a reproducible workflow, retaining physics related quality measures and thereby successfully integrating new compression schemes. Moreover, the jump in compression ratio is remarkable and clearly motivates further studies if and how such high compression ratios can be obtained with a high-bandwidth implementation of grok/j2k.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Achieving acceptance in the user community

While lossless compression is well-accepted by end users of LEAPS facilities, the introduction of lossy compression scheme can be much more difficult, as the user community needs first a solid guarantee that the compression does not affect the quality of the interpretable data. The tomography community stands apart in this aspect, as the JPEG2000 compression has been validated more than 20 years ago, and notably included in the DICOM exchange format[12] for medical imaging.

For other techniques, the choice to either use lossy compression or keep only reduced data (e.g. for powder diffraction or small-angle scattering) remains difficult as it removes the ability to reprocess completely a dataset-which is useful not only in case some parameters were not optimal, but also for the open science credibility which is provided by the availability of the raw data.

It should be noted that the choice to use lossy compression may be resource-driven, due to the costs of data storage, or to the difficulty of transferring the data at a speed of 10s of GBytes/s from an increasing number of detectors in a given facility.

Finally, the role of scientific bodies in addressing these issues is very important, to define acceptable data standards collectively (see the example of DICOM above, or the work of the IUCr for crystallography data standards).

[12] https://www.dicomstandard.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Compression in X-ray diffraction experiments; serial crystallography as a leading example

In this report, we have focused on data compression in imaging experiments, since this is an area where most data at our facilities is currently uncompressed. In X-ray diffraction, data compression is already a common practice, but there is still strong demand for improving compression. We present a survey of compression in serial crystallography as an example of cutting-edge practice in this field.

In serial crystallography, X-ray diffraction patterns are obtained from a large number of protein crystals, each typically with a random orientation. By indexing each pattern to find the crystal orientation, these patterns can be merged to obtain a full set of Bragg peak intensities. This makes it possible to study proteins that do not form large, high-quality crystals, such as membrane proteins, and also enables time-resolved experiments where a change is initiated in the protein (e.g. by light) before probing the sample with X-rays after a variable time delay. This technique was originally developed at free-electron lasers (FELs), where the ultrashort X-ray pulses can "outrun" radiation damage effects to obtain better-quality data (Doerr 2011, doi.org/10.1038/nmeth0411-283). However, there is also increasing use of serial crystallography at synchrotrons (Leonarski et al 2023, doi.org/10.1107/S1600577522010268).

This technique relies on acquiring large numbers of diffraction patterns; with modern detectors, it is possible to obtain thousands of patterns per second, and uncompressed data sizes can reach 1 PB per day or more. Currently, a range of techniques are used for data reduction.

Lossless image compression

At synchrotrons, protein crystallography experiments are primarily done with photon-counting detectors, which provide an effectively-noise-free count of the number of photons arriving in each pixel. Generally, the statistics of pixel values in these experiments are highly non-uniform, with Bragg peaks having high intensities, but with many pixels containing few or zero photons. A variety of lossless compression algorithms can achieve a reasonable compression ratio - of order 5 is reported in Galchenkova et al. 2024 (doi.org/10.1107%2FS205225252400054X). Typically, pre-processing with the Bitshuffle algorithm is used; in regions of the image with low intensities, the bitshuffled data contains long sequences of zero-valued bytes that are easily compressed.

At FELs, photon counting detectors cannot be used because all the photons arrive simultaneously, and so integrating detectors are used which measure the total energy deposited in each pixel. This measurement has some noise, and the resulting images do not losslessly compress well. However, typically the noise is significantly less than one photon, and so "photonization" can be applied to the image - rounding pixel values to the nearest photon, or some fixed fraction of a photon - to effectively reduce this noise and improve compressibility, albeit at the cost of additional processing.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Lossy image compression

As data rates in these experiments increase, there is demand for higher compression ratios that can only be achieved by lossy compression. As with all lossy compression methods, it is important to evaluate the effects of compression on data quality, by repeating the data analysis with compressed data and assessing the quality of the result. Quality metrics are discussed later in this section.

Galchenkova et al. 2024 investigates a variety of approaches such as binning pixels or quantizing images before applying lossless compression. One intriguing result is that non-uniform quantization, where each pixel value is converted to 3 or so leading bit values plus an exponent, has little effect on data quality while achieving compression ratios of 30 or more. (When an image is quantized this way, weak diffraction spots at large angles which encode high-resolution structure will not lose precision.)

Ultimately, the useful information in the diffraction consists of Bragg peak intensities, so one proposed approach is to only store pixel values in the vicinity of the Bragg peaks. However, in serial crystallography, weak peaks at high scattering distance (*Q*) may not be reliably detectable in individual images by peak finding algorithms, but emerge when combining many patterns. So, attempting to store only data in regions of Bragg spots found in initial peak finding (prior to indexing) reduces data quality. However, it is reported that a hybrid approach, ROIBIN-SZ, where lossless compression is used in regions of interest around detected Bragg peaks and lossy compression elsewhere, does a better job of preserving data quality which achieving a compression ratio of around 40 (Underwood et al. 2023, doi.org/10.1080/08940886.2023.2245722).

A similar approach has been developed at ESRF for on-the-fly lossy data compression using a statistical analysis (sigma-clipping) of serial crystallography images[13] [14], making it possible to significantly reduce the data volume, using a single GPU for 250 Hz data processing of 4 Mpixels data streams.

Image triage

In serial crystallography, there can be a large fraction of images where no diffraction is observed; in particular, in liquid jet experiments at FELs, most X-ray pulses do not hit a crystal and the hit rate can be of order 1% or lower. The first step of indexing a diffraction pattern is to find the Bragg peaks in the image, and it is common to reject not only images with zero detectable Bragg peaks, but also those with few peaks. Typically, this rejection process is applied offline, to data saved on disk, but live rejection of bad images is being developed.

The number of detectable Bragg peaks will depend on the quality of the diffraction pattern; a partial hit on a low quality crystal will produce a more limited number of visible Bragg peaks at small angles. As a result, data quality is relatively robust across different cutoff criteria, since diffraction patterns with borderline quality will contribute little to the quality of the final merged dataset. However, it is

[13] https://indico.psi.ch/event/12738/contributions/38902/attachments/22756/40116/2022-09-21-NoBugs2022-short.pdf

[14] https://journals.iucr.org/a/issues/2022/a2/00/a61396/a61396.pdf

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

challenging to perform Bragg peak detection sufficiently quickly to reject images during data acquisition. Approaches for faster Bragg peak detection include developing faster algorithms (Hadian-Jazi et al. 2021, doi.org/10.1107/S1600576721007317) and implementing them on accelerators such as GPUs (Ponsard et al. 2020, doi.org/10.1107/S1600577520008140).

Online data processing

After identifying good diffraction images, the next step in data analysis is to perform indexing, to find their orientation, and then to merge all the indexed images into a single 3D dataset, from which the intensity of each reflection can be extracted. This 3D dataset is much smaller in size than the set of diffraction images, since at each possible orientation many images are combined. So, performing this data processing live can reduce the costs of storage, for example by allowing the original images to be migrated to cheaper storage relatively soon, or even by removing the need to save them.

The main challenge is to perform the analysis, particularly indexing, sufficiently fast to match the rate of data collection. A recent implementation of indexing, using the PyTorch machine learning framework, can run on both CPU and GPU, achieving similar performance to existing algorithms in terms of quality metrics, and a speed-up of order x 7 when running on CPU and x 1000 when running on a high-performance GPU (Gasparotto et al, preprint at doi.org/10.26434/chemrxiv-2023-wnm9n).

Quality metrics

Different metrics can be applied to different stages of the analysis, as discussed for example in Karplus and Diederichs 2015 (doi.org/10.1016/j.sbi.2015.07.003).

Broadly speaking, these metrics can be divided into the following categories, and it is worthwhile to evaluate the effects of lossy compression and processing algorithms using a combination of these:

- Metrics of data precision and reproducibility. For example, the two halves of a merged dataset should be theoretically identical, and $CC_{1/2}$ measures the similarity between the two halves using a Pearson correlation coefficient (the higher the better).
- Model refinement quality, i.e. the agreement between the final molecular model and the data. This is found by calculating the peak intensities that would be expected, given the final molecular model, and comparing them to the measured values. In the metric R_{free} a fraction of the experimental data is used only for making this comparison and not for model building, to reduce overfitting. (Lower R_{free} values correspond to better agreement.)
- Metrics can also be applied to anomalous signal; for example, CC_{anom} is the correlation coefficient between the anomalous differences.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.

Conclusion

Due to the increasing data rates in photon science experiments, data reduction is increasingly crucial for limiting the costs of data storage and enabling new, data-intensive experimental techniques. As well as developing experiment-specific data reduction methods, there is a lot of potential to benefit from existing compression algorithms and approaches. Machine learning techniques are a particular area of growth; as demonstrated here, de-noising can make data more compressible, and machine learning can also be used to optimise compression parameters. In addition to file compression, data can also be reduced for example by rejecting bad data or performing online data analysis and deleting the raw data entirely.

The performance of a data reduction method will depend strongly on the data itself, so testing is important. In particular, methods such as lossy compression may affect the final results of analysis. So, it is important to run the full data processing pipeline on the data after compression and apply appropriate metrics to judge whether the scientific quality is affected. Even with similar techniques, the performance may vary for different beamlines or samples, and so it is important to re-evaluate this for new datasets. This is also important for building trust in a compression technique among users. As a result, it is very useful to develop reproducible workflows for processing data and testing compression, for example using workflow engines; this also of course has benefits for transferring data compression methods from one facility to another.

Once a data reduction method has been proven to work, it is important to ensure that the data reduction is performed in a way that ensures interoperability; the data reduction should not make it more difficult for the user to analyse the data, or limit the analysis tools they can use. At our facilities, HDF5 is a widely-used format that makes it possible to store large, structured datasets along with metadata. By integrating new compression methods into HDF5, the user can work with the data without changing their analysis routines. In this project, we have demonstrated the integration of JPEG2000 codecs for imaging into HDF5, by incorporating them into the Blosc2 meta-compression library.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 871072.