

Massive compression for high data rate macromolecular crystallography (HDRMX): impact on diffraction data and subsequent structural analysis

Herbert J. Bernstein,^{a*} Alexei S. Soares,^{b*} Kimberly Horvat^c and Jean Jakoncic^b

Received 15 April 2024

Accepted 16 January 2025

Edited by M. Yamamoto, RIKEN SPring-8 Center, Japan

This article is part of a collection of articles from the IUCr 2023 Congress in Melbourne, Australia, and commemorates the 75th anniversary of the IUCr.

Keywords: compression; lossy compression; data statistics; experimental phasing; omit map; structure refinement; Bragg reflections.

Supporting information: this article has supporting information at journals.iucr.org/s

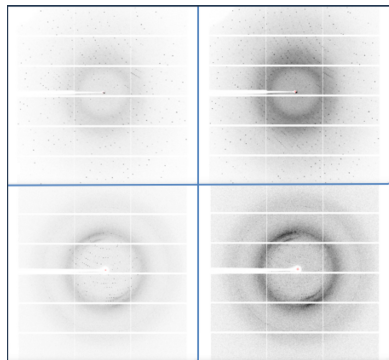
^aFresh Pond Research Institute, c/o NSLS-II, Brookhaven National Laboratory, Bldg 745, Upton, NY 11973-5000, USA, ^bNational Synchrotron Light Source II, Brookhaven National Laboratory, Bldg 745, Upton, NY 11973-5000, USA, and ^cStony Brook University, 100 Nicolls Rd, Stony Brook, NY 11794, USA. *Correspondence e-mail: hbernstein@bnl.gov, soares@bnl.gov

New higher-count-rate, integrating, large-area X-ray detectors with framing rates as high as 17400 images per second are beginning to be available. These will soon be used for specialized macromolecular crystallography experiments but will require optimal lossy compression algorithms to enable systems to keep up with data throughput. Some information may be lost. Can we minimize this loss with acceptable impact on structural information? To explore this question, we have considered several approaches: summing short sequences of images, binning to create the effect of larger pixels, use of JPEG-2000 lossy wavelet-based compression, and use of Hcompress, which is a Haar-wavelet-based lossy compression borrowed from astronomy. We also explore the effect of the combination of summing, binning, and Hcompress or JPEG-2000. In each of these last two methods one can specify approximately how much one wants the result to be compressed from the starting file size. These provide particularly effective lossy compressions that retain essential information for structure solution from Bragg reflections.

1. Introduction

This paper continues a discussion of approaches to compression of diffraction images in macromolecular crystallography (MX) that began with consideration of optimal lossless compression (Bernstein & Jakoncic, 2024). In this paper we consider the increasing need for lossy compression, arising as data rates increase in MX and as data retention requirements also increase. Lossy compressions have been in use for a long time, even in crystallography. Reducing a diffraction data set to a list of structure factors is, indeed, a dramatic lossy compression. With new detectors and many beamline upgrades on the horizon, one will need to use lossy compression to keep up.

The general concept of making 'raw' data available for all new PDB structures deposited, as part of improved FAIR [FAIR data meet principles of Findability, Accessibility, Interoperability, and Reusability (FAIR)] principles, may also drive the use of lossy compression. Lossy compression, with the goal of minimal structural information loss, is a viable method to archive diffraction data sets permanently. Here we investigate a few of the most promising available methods to apply lossy compression on two test samples, one used for sulfur single-wavelength anomalous diffraction (S_{SAD}) phasing, and one for molecular replacement. We ran tests using various parameters to determine how far we could push



OPEN ACCESS

Published under a CC BY 4.0 licence

the compression ratios and still determine the same structure. It is important to note that, with this approach, the diffraction data that will be stored long-term are exactly the data that were used to solve the structure.

1.1. High-data-rate MX

Enabled by changes in technology, MX increasingly is able to extend its focus from the average state observed in a single crystal, or from merged data from a few crystals, to studies of families of distinct structural states arising from data collection from many crystals. This gradual transition, enabled by brighter sources, smaller beams, and new detectors, is driving a series of disruptive changes in the way diffraction data are collected, processed, and archived. Hardware improvements, such as fast, high-resolution, low-noise and high-dynamic-range detectors, high-brilliance X-ray micro-beams, and automated sample handling combined with autonomous operation, are generating high-data-rate and high-data-volume data streams. Conventional software packages and pipelines, designed for simpler single-crystal experiments and one-node serial processing, cannot support or even keep up with such rates and volumes. Networks and computational resources have had to be regularly upgraded. Bottlenecks in data-processing pipelines have been removed, sometimes by converting serial execution to parallel execution on multiple nodes. But those changes themselves have generated yet more network and computational load.

Higher flux, smaller beams, and faster detectors open the door to experiments with very large numbers of very small samples that can reveal polymorphs and time-resolved structural information on dynamic effects, but require re-engineering of approaches to the storing of images. Management of orders-of-magnitude more images and limitations of file systems have favoured a transition from simple one-file-per-image systems such as CBF (Bernstein *et al.*, 2020) to image-container systems such as HDF5. This further increased the load on computers and networks, and the use of data coming from multiple runs at multiple beamlines has required a re-examination of the presentation of metadata. Over the past few years the high-data-rate MX community has recognized the importance of complete and consistent metadata. One requires this metadata so data sets can be easily processed at any site from data collected at different times, at another facility, or at multiple facilities, or months and years in the past (Bernstein *et al.*, 2020). Full metadata associated with the diffraction data set further makes FAIR principles attainable now.

When we began thinking about this, data rates were typically 100 frames per second (fps) and were approaching 1000 fps in some cases. Now we are facing data rates exceeding 15000 fps (Hiraki *et al.*, 2021). We must either discard a significant portion of the data collected or make serious use of greater compression ratios. Indeed, in both X-ray free-electron laser serial crystallography and synchrotron serial crystallography, it has long been acceptable to discard frames that appear to contain no significant number of reflections.

1.1.1. Terminology of compression ratios

For a review of the general terminology of compression, see Lelewer & Hirschberg (1987). In this paper we will follow the information-theory approach of Shannon (1948) in which, for this crystallographic application, we focus on the information in diffraction images as streams of numbers representing the intensities of pixels. We define a ‘compression ratio’ as the quotient of the size of the original image to the size of the hopefully smaller compressed image that will be processed or stored. For this to make intuitive sense, one must use comparable units for both the numerator and the divisor, *i.e.* bytes over bytes or bits over bits. For example, for the Dectris Pilatus detector as released in 2007 (Bernstein, 2010), the original image data as collected consisted of four-byte-long integer pixels, but the data as received had gone through a faithful four-to-one byte-offset compression ratio using the CBF library, *i.e.* yielding a compression ratio of four-to-one.

Different compression algorithms can achieve different compression ratios for different image data, but if we wish to retain all the information in an image, as Shannon proved, there is an inherent limit to the highest faithful compression ratio that can be achieved, called the ‘entropy-limit’. Some of the information in the stream may be noise or may be irrelevant to the experiment, and we may choose to ‘lose’ part of the information, giving us a higher compression ratio beyond the entropy limit.

In this paper we focus on the extra compression ratio (ECR) achieved beyond the entropy limit. As a practical matter, for most diffraction images, the entropy limit is within a factor of two, or less, of the simple byte-offset compression ratio, so we use the byte-offset compression as our base for computing the ECR. Other compression reference points are possible, but this one is a realistic approximation within a factor of two or less to that with which most macromolecular beamlines currently work.

1.2. Lossless compression

When crystallographic diffraction images are recorded, the common practice has been to retain all details of what has been recorded. Such images have been compressed by lossless compression so that all details are recovered when the images are decompressed. For a general review of lossless compression, see Rahman & Hamada (2019). Bernstein & Jakoncic (2024) discussed how the algorithms related to the lossless Lempel–Ziv–Welch compression algorithm (LZW) (Ziv & Lempel, 1977; Welch, 1984) are, in many cases, the best choice for a wide range of applications. However, LZW and its derivative algorithms were not implemented for MX data because of an active patent (Welch, 1985) that would have resulted in significant added costs. As a result, for decades, the MX field employed other algorithms (Abrahams, 1993; Bernstein *et al.*, 1999; Ellis & Bernstein, 2006). The Dectris Pilatus 6M (Brönnimann *et al.*, 2003; Brönnimann, 2005) adopted the CBF format with byte-offset compression (Bernstein, 2010). By 2012, the LZW patents had expired and available computer hardware could handle higher data-

framing rates and more complex compressions. In 2013 Dectris designated Collet's LZ4 compression (Collet, 2011), a very fast byte-oriented version of Lempel–Ziv compression without an entropy stage, as the compressor for their Eiger detectors (Donath *et al.*, 2013). Later the compression ratios were improved by adding a pre-conditioning stage using Masui's bit-shuffle algorithm (Masui *et al.*, 2015). In the interim, others have developed and experimented with further LZW compression algorithms. A popular, fast, and efficient algorithm is Z-standard ('zstd') (Collet & Kucherawy, 2021).

1.3. Lossy compression

As noted in Kroon-Batenburg & Helliwell (2014), 'we have to consider the possibility of lossy compression'. Holton applied a lossy-compression algorithm to ADSC¹ images from a lysozyme crystal by splitting them into images that contain reflections and those that contain the background (see https://bl831.als.lbl.gov/~jamesh/lossy_compression). The background images are then compressed so the noise level (*i.e.* the variance of the background) stays roughly the same, and finally the two types of images are recombined. He achieved a compression ratio of 34:1 (the original size of the data to the compressed size) without changing the visual perception of the images and without changing the derived structure factors significantly [$\langle \Delta F/\sigma(F) \rangle = 0.6$]. Ferrer *et al.* (1998) examined lossy data compression of CCD images with discrete cosine or wavelet transforms, achieving approximately a 30:1 compression ratio. They also found little effect on integrated intensities except for weak reflections, which may be related to altered statistical noise in the background or to the algorithm flattening the weak intensities.

Here, and from now on in this paper, comparisons of effect are between diffraction-data images integrated in the best possible way to give intensities, and intensities integrated in exactly the same way from images that have been compressed and then decompressed (see Section 2.3).

However, the obvious advantages of lossy compression for network transfer and disk space could be outweighed by the (unforeseen) future need for raw images. One consideration could be the diffuse scattering of less-ordered crystals that occurs in areas that we normally designate background and which is weak and varies slowly. Currently, no one has established the extent to which such compression algorithms will significantly affect diffuse intensities in reciprocal space. Underwood *et al.* (2023) have proposed combining lossless compression in rectangles around Bragg peaks with lossy compression for the remaining background to cope with the coming data volume loads for serial crystallography. They call their algorithm ROIBIN-SZ. Galchenkova *et al.* (2024) considered use of lossy compression in serial crystallography and saw encouraging results with 'real time hit finding, binning, quantization to photons, and non-linear reduction of the dynamical range', but noted a significant loss of resolution

achieved with ROIBIN-SZ in comparison with techniques that preserve the background.

Fields such as astronomy that must deal with very large volumes of mostly empty image data have long employed lossy compression; a popular example of useful software is Hcompress (White *et al.*, 1992). Hcompress is a method which is based on the Haar-wavelet-based Haar transformation (Fritze *et al.*, 1977). The Haar wavelet is a flat-topped wavelet with discontinuous transitions that help in modelling sharp peaks such as stars and potentially diffraction-peak images. The purpose of Hcompress is to compress astronomical images quickly, reducing the redundancy of pixels displaying bright celestial object images. In addition to that, astronomical images contain a lot of noise in the dark-sky background, which can be removed. These characteristics of astronomical images are shared by MX diffraction images.

As we will see, our experiments bear out the value this suggests in using Hcompress for MX. Applying these approaches can result in significant cost savings (see Section S1 in the supporting information). If we assume a data-management policy requiring very long term storage of data, but that the data can move to lower-cost off-line storage a decade after collection, the current annual cost estimate for archival lossless storage at USD 14400 per annum per beamline can be converted to a steady state cost of USD 144 K (for the current year's data plus nine previous years) per annum per beamline. This could be reduced to less than USD 1000 per annum per beamline if lossy compressions with ECRs averaging more than 144:1 prove satisfactory.

For all examined lossy compression schemes, the computational resources devoted to data compression are small compared with existing needs such as data reduction and image scoring. Furthermore, these computational resources can be deployed during periods when data acquisition is not occurring, such as scheduled equipment maintenance periods. Hence, we believe that the computational overhead needed for compression can be negligible. Decompression will be built into data processing plugins, essentially resulting in no decompression penalty for all users.

1.4. The FAIR principle and the US White House mandate

There are two factors in tension with any data-management approach that discards any diffraction images or even any portions of diffraction images. There has been a world-wide movement of science towards the so-called 'FAIR' principles (Wilkinson *et al.*, 2016), which require that scientific data be Findable, Accessible, Interoperable, and Reusable. In 2022, the USA formally joined this movement with the White House mandate that data from Federally funded research would become freely available by the end of 2025 (White House, 2022). Clearly, we must devise some rational structure to achieve an appropriate balance between the desire to share all data and the availability of the necessary resources.

Making the right choices in lossy compression should help to achieve that appropriate balance. For MX there are some special issues to consider in achieving the necessary balance.

¹ Area Detector Systems Corp produced a popular and very successful line of charge-coupled-device-based area detectors.

In some protein dynamics studies one needs to keep all diffraction data from all samples with sufficient data to show the aspects of dynamics supported by non-Bragg scattering. One must also reveal aspects supported by careful clustering of families of multiple states, shown in appropriately chosen Bragg diffraction images for which most of the background pixels can be discarded. In addition, for non-dynamic studies, most images do not provide useful data, and most pixels in the images that do provide useful data are not relevant to the final results. It should be sufficient in the course of applying lossy compression to retain an appropriately chosen subset of between one-half and one percent of the images as lossless compression images, in order to detect the cases in which non-Bragg background data are relevant and scientifically interesting.

Here, we focus our attention on a few types of lossy compression: frame summing, pixel binning, JPEG-2000, and Hcompress, and optimal combinations of these. We report the effects of these compressions on analysis and structural refinement for two complete representative test data sets. We have found a few promising lossy compressions that can be used routinely, depending on the type of experiment planned or on the intensities of the observed reflections.

Future steps are to investigate the impact of lossy compression choices on real-life data collections by making lossy compression options available, first as redundant parallel paths in data-processing production pipelines, and then, if justified by the results, as selectable alternative processing choices. This builds on Holton's approach of segregating the background processing from reflection processing.

2. Materials and methods

2.1. Data collection

Data were collected at the National Synchrotron Light Source II (NSLS-II) Highly Automated Macromolecular Crystallography (AMX) beamline 17-ID-1 (Schneider *et al.*, 2022) at Brookhaven National Laboratory, using an Eiger X 9M detector (Förster *et al.*, 2016). We collected two distinct data sets: (1) a data set from an HIV reverse transcriptase crystal (Hollander *et al.*, 2023; Prucha *et al.*, 2023; Chan *et al.*, 2020; Duong *et al.*, 2020) at 13.5 keV for molecular replacement studies, and (2) a redundant data set on a lysozyme crystal at 7.5 keV to increase the anomalous scattering factor of sulfur for S₂ SAD phasing. There were a total of 1200 and 1800 rotation frames of 0.2° each for the HIV reverse transcriptase and the low energy S₂ SAD lysozyme samples, respectively.

2.2. Lossy compression methods tested

2.2.1. Image summing

When data have been collected using fine φ -slicing (0.2° frame⁻¹), a very effective compression is to sum a small number of sequential frames, pixel-by-pixel. For this series of experiments, we achieved this summing with the program *merge2cbf* from the *XDS* package (<https://xds.mr.mpg.de/>

[html_doc/merge2cbf_program.html](https://xds.mr.mpg.de/html_doc/merge2cbf_program.html)). For the HIV reverse transcriptase we tried summing by 2's, by 5's, and by 10's, and experimented with summing by 20's and 40's, essentially converting the angle of each slice to an angle two, five, or ten times as coarse, as well as even coarser slices (20 and 40). Recall that, in the 1980s, employing X-ray film as detector, we might have taken one-to-five-degree images [here the motive was to take as wide an image as possible (fewer films to process) while avoiding collision of spots on the film: large unit cells, smaller increments], the equivalent of 5–25 of these small frames. We label n -fold coarse slicing 'SUM n '. For moderate values of n , this has very little effect on the quality of the peaks collected other than to make it less likely to encounter 'partial reflections', but it does cause a linear increase in the level of the background, making it more challenging to detect weak peaks.

2.2.2. Pixel binning

Another very effective compression is to sum pixels in the plane of the image which we call 'binning'. For example, in 2 × 2-fold binning, each even-numbered pixel in each even-numbered row is replaced by the sum of that pixel with the pixel to the right and the pixel above and the pixel diagonally above and to the right. We label that two-fold binning 'BIN2', and in general label n -fold binning 'BIN n '. This has the effect of reducing the number of pixels by a factor of four, or equivalently of increasing the size of each pixel by a factor of two in each direction. We experimented with BIN2, BIN4, BIN6, BIN8, and BIN10 for the reverse transcriptase. For the lysozyme data set, we only applied BIN2.

By combining the coarsening of the slicing with binning by two, one can achieve a compression by a factor of eight (SUM2 + BIN2) or twenty (SUM5 + BIN2) at very low cost in terms of both time and statistics. For these experiments the binning was done by use of the program *cif2cbf* from the *CBFlib* package (<https://github.com/yayahjb/cbflib>). The use of modest amounts of coarsening of the slicing with binning combines very well with any of the popular lossless compressions including those from *CBFlib* as well as with *bslz4* and *bszstd*.

2.2.3. JPEG-2000 and Hcompress

Two compression techniques from other scientific fields may be useful for diffraction images: JPEG-2000 from the image-processing industry (Santa-Cruz *et al.*, 2002) as released in the *OpenJPEG* package (ISPGGroup, 2023), and Hcompress from astronomy (White, 2019). JPEG-2000 uses the Daubechies wavelet, which is a smoother generalization of the Haar wavelet, which is now referred to as DB-1.

For these studies, JPEG-2000 requires a preliminary conversion of images to 16-bit greyscale tiff format (Data, 1947; Poynton, 1992), while Hcompress requires a preliminary conversion to astronomical fits format (Pence *et al.*, 2010) after conversion to tiff. We used the *siril* package for conversion from tiff to fits (<https://siril.org/>). (For our data, use of 8-bit greyscale tiff proved to be too aggressive a lossy compression

compared with 16-bit tiff, though it might be suitable for other data with less dynamic range and a flatter background.)

For both samples, we tried various compression levels available in the JPEG-2000 and Hcompress applications.

Finally, to study higher compression ratios, we tried various combinations of summing, binning, JPEG-2000, and Hcompress. In these the combinations of summing pairs of images, 2×2 binning and Hcompress were particularly interesting in terms of high compression ratios with good data processing statistics. We label JPEG-2000 compression with a target compression ratio of n -to-1 as 'J2K n '. We label Hcompress compression with m -fold scaling as 'HCOMP m '.

We tried various combinations of compressions on a lysozyme data set of 1800 frames and on an HIV reverse transcriptase data set of 1200 frames, both collected at $0.2^\circ \text{ frame}^{-1}$, for 360° of lysozyme and 240° of HIV reverse transcriptase. Examples of uncompressed native lysozyme, BIN2_SUM2 (eight-to-one compressed) lysozyme, uncompressed native HIV reverse transcriptase, and BIN2_SUM2 (eight-to-one compressed) HIV reverse transcriptase are shown in Figs. 1(a), 1(b), 1(c), and 1(d), respectively. Note that in each case the compressed images preserve almost all significant reflections against backgrounds stronger than in the native images. Notice as well that the two native images differ significantly in terms of density and intensity of peaks across the images and uniformity of the backgrounds.

All 1800 original diffraction frames for the S_SAD lysozyme, are available on zenodo ([https://zenodo.org/records/](https://zenodo.org/records/12701763)

12701763) with all compressions applied and the data back-converted to CBF format.

2.3. Data processing and analysis

The two original data sets (HDF5 files) were analysed with the *fast_dp* pipeline (Winter & McAuley, 2011). An XDS.INP file was derived from the intermediate XDS.INP files that *fast_dp* creates for spot finding, indexing and integration. Execution of XDS generated XDS_ASCII.HKL files which contained corrected intensities. These were processed further using *POINTLESS* and *AIMLESS* (Evans & Murshudov, 2013) for space-group selection and for scaling. We then used the two resulting MTZ files, one for the reverse transcriptase and one for the lysozyme, for structure solution.

The reverse transcriptase structure was refined using PDB (Bernstein *et al.*, 1977; Berman *et al.*, 2003) entry 8STS (Hollander *et al.*, 2023) as a starting model for the molecular refinement. The refined model was manually corrected to account for a loop/domain motion using *COOT* (Emsley & Cowtan, 2004). The final model was refined using *REFMAC* (Murshudov *et al.*, 1997; Vagin *et al.*, 2004) and *COOT*.

The lysozyme uncompressed data were processed using the same manual workflow consisting of XDS, *POINTLESS*, and *AIMLESS*. The reduced file was used for a SAD-phasing experiment using the *hkl2map* (Pape & Schneider, 2004) and the *SHELXC*, *SHELXD*, and *SHELXE* components of the *SHELX* system (Sheldrick, 2008) software packages. We selected these packages since they offer manual tuning of some parameters. The output file was used for further processing using density modification (DM) (Cowtan, 2010) and model building from *ARP/wARP* (Langer *et al.*, 2008) using a semi-automated workflow.

All lysozyme and HIV reverse transcriptase compressed data were reduced using the corresponding XDS.INP file with a few modifications to account for binning and summing when required. Pixel sizes, number of pixels, beam centre, and detector gaps, plus oscillation range and number of frames, were updated accordingly. No additional modifications to XDS.INP were required for the JPEG-2000 and HCOMP compression *per se*.

Subsequent models from the reverse transcriptase compressed data were refined using the original model as a starting point for rigid-body and atomic refinement using *REFMAC*. All lysozyme compressed reduced data were used following the same workflow employed to solve the uncompressed S_SAD.

3. Results

3.1. Achieved compression ratios

As one might expect from the differences in the native images in both data sets [Figs. 1(a) and 1(c)], of the extra compression ratios achieved from the two data sets, as shown in Figs. 2 and 3, the lysozyme data set is a little more difficult to compress than the HIV reverse transcriptase data set. That is, there is clearly more information in image 1(a) than 1(c).

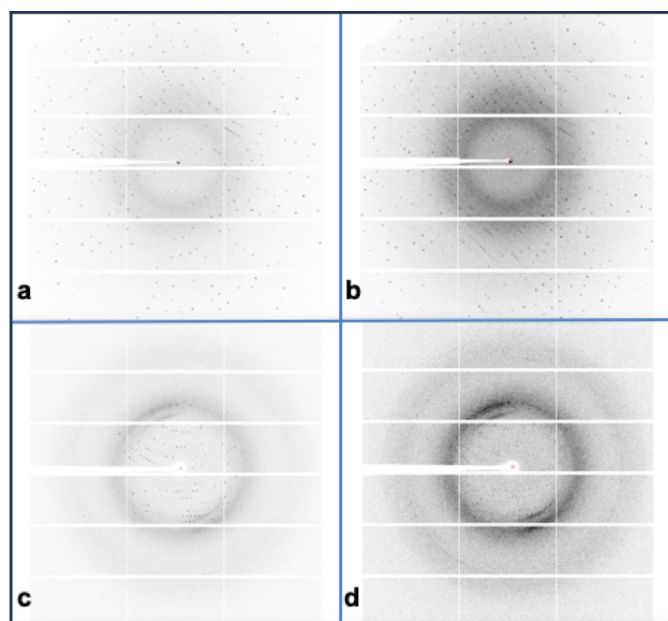


Figure 1

(a) Lysozyme frame 900 of 1800 from the native data set. (b) Lysozyme frame 450 of 900 from the 8:1 compressed BIN2_SUM2 data set, added to the one after it, and then with pixels summed 2×2 . Notice the much stronger background with well preserved peaks over the entire image, and that one can pick out a few more peaks in (b) than in (a). (c) HIV reverse transcriptase frame 600 of 1200 from the native data set. (d) HIV reverse transcriptase frame 300 of 600 from the 8:1 compressed BIN2_SUM2 data set. Notice the much stronger background in (d) than in (c).

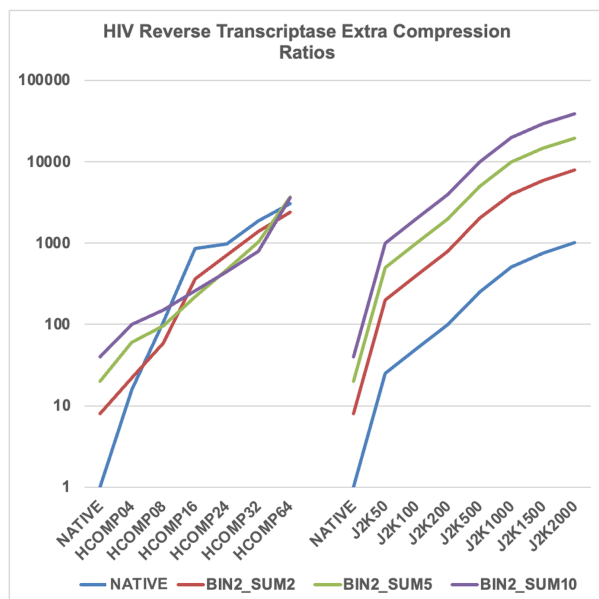


Figure 2

Wide selection of available lossy compressions for HIV reverse transcriptase. Note that many of the compression ratios exceed 500:1. Not all compressions tested are included in this figure. A table with all compressions is included in the supporting information.

Many of the available extra compression ratios tested exceeded 500:1. An appropriate subset of these compressions was tested for their impact on the effectiveness and quality of structure solution.

We are surveying the impact of lossy compressions on data reduction as well as on the ability to solve 3-D crystal structures. One can gauge lossy compression effects in several ways: (1) first looking at the corresponding overall diffraction

pattern, (2) then observing impacts on weak and strong reflections, observing the diffraction image of the integrated reflection intensity divided by surrounding background, by looking at data reduction key parameters, and (3) finally looking at the refined structure analysis and the corresponding electron density maps.

3.2. Compression results for lysozyme

The native and compressed data for lysozyme were processed with *XDS* (Kabsch, 2010); *AIMLESS*; *POINTLESS*; *SHELXC*, *SHELXD*, and *SHELXE* components of the *SHELX* system (Sheldrick, 2008), and other CCP4 components (Evans, 2014).

Fig. 4 shows the data reduction statistics and extra compression ratios for lysozyme data sets using various compression methods. The corresponding Tables S4 and S5 in the supporting information highlight the most commonly used parameters [R_{pim} (precision indicating merging R , scaled to give precision), completeness, mean $I/\sigma(I)$, unique reflections] as well as a subsequent set of parameters to illustrate potential damage to data statistics prior to S-SAD phasing (anomalous correlation and mid-slope anomalous probability). The objective is to obtain the maximum compression that is consistent with keeping R_{pim} low and both $I/\sigma(I)$ and completeness high. Numbers for the highest resolution shell (1.67–1.64 Å) are shown there in parentheses. For simplification, other parameters usually displayed are not included in Tables S4 and S5. 1800 frames were collected at 7.5 keV to improve the S-SAD signal to maximize the likelihood of successful phasing using anomalous signal from the ten S atoms present in the lysozyme. Note that the protein was crystallized in an NaCl solution resulting in the observation of seven additional peaks.

Here we are investigating the effects of binning pixels, summing frames, JPEG-2000, and Hcompress lossy compressions, and subsets of combinations. The AMX beamline relies on an EIGER X 9M detector with a minimum distance of 100 mm between the sample and the detector, limiting the highest achievable resolution for a given wavelength. This results in an overall completeness of 88.9% in the 20.00–1.64 Å range.

Note that after we first attempted lossy compression on the HIV reverse transcriptase we concluded that some of the compressions tested (BIN4, BIN6, and higher; various summing sets) were not likely to be useful in the lysozyme case. Therefore, for lysozyme we focused on Hcompress, JPEG-2000, and the combinatorial effects of BIN2_SUM2, BIN2_SUM5, BIN2_SUM10 with either JPEG-2000 or Hcompress.

Fig. 4 and Tables S4 and S5 highlight the most important parameters selected to probe the effects of lossy compression when compared with the ‘uncompressed’ data. From the data-reduction statistics perspective only, one can see clearly that all compressions tested have significant effects on the data statistics. These effects increase with increased compression by JPEG-2000 and Hcompress.

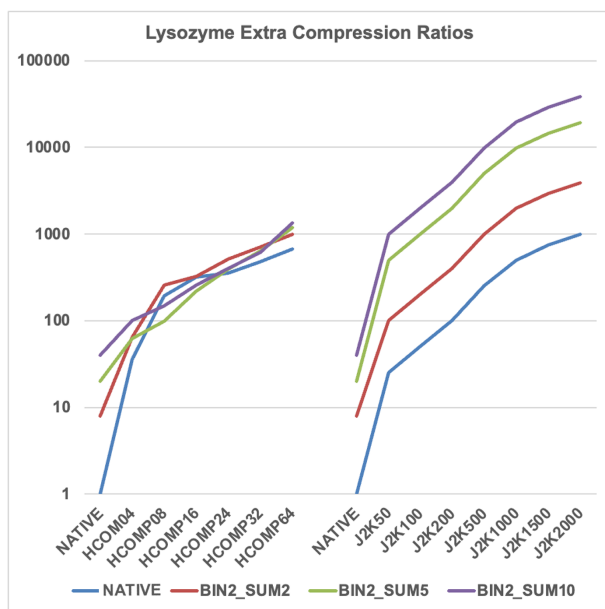


Figure 3

Wide selection of available lossy compressions for lysozyme. Note that many of the compression ratios exceed 500:1. Not all compressions tested are included in this figure.

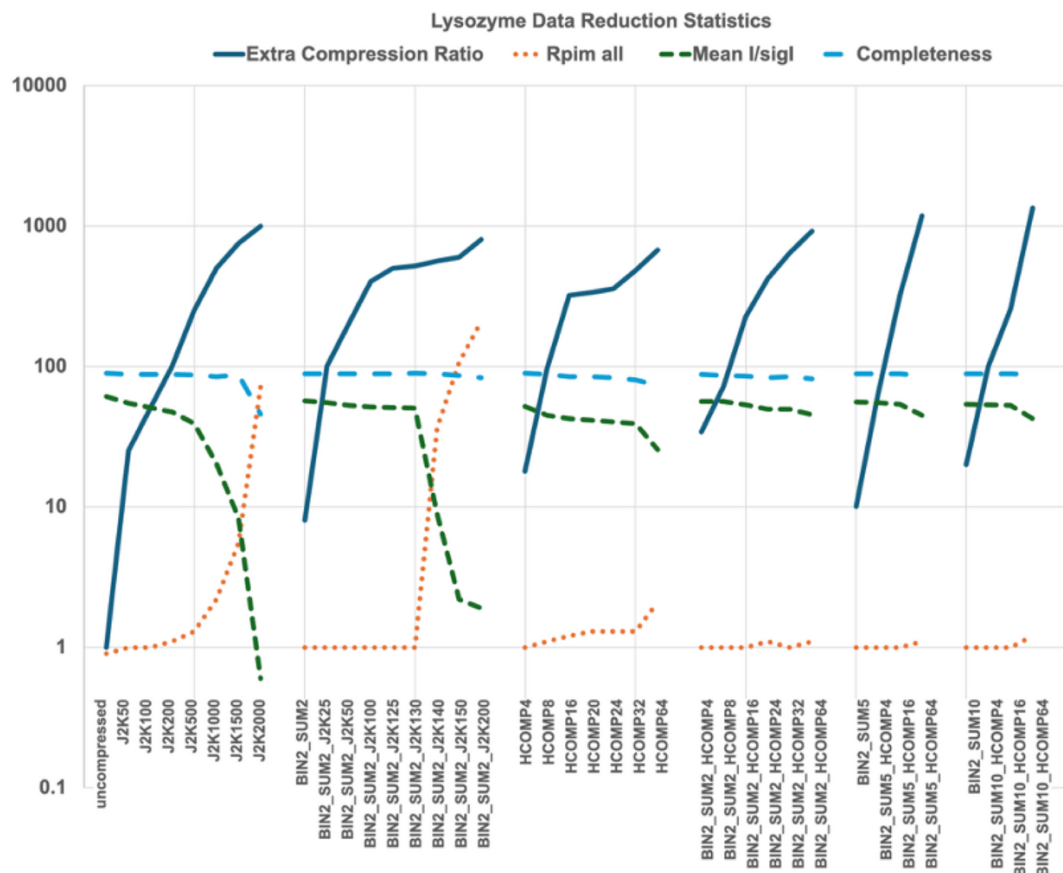


Figure 4

A portion of the data reduction statistics for all lysozyme data sets. The figure shows the most commonly used parameters, R_{pim} , completeness and mean $I/\sigma(I)$. The vertical axis is the numeric value for the extra compression ratio achieved as well as for each of the commonly used parameters. More complete information is given in the supporting information.

3.2.1. JPEG compression

For lysozyme, we tested JPEG-2000 compression levels ranging from 50:1 to 2000:1; measured compression ratios reflected the ratio requests, *i.e.* the extra compression ratios were approximately half the requested J2K overall compression ratios. Applying increasing JPEG-2000 compression levels resulted in gradual degradation of $I/\sigma(I)$ and R -factors, with J2K2000 data being severely damaged; one might even say destroyed (see the two left panels in Fig. 4). Completeness of J2K2000 data was half that of the uncompressed data, and R -factors as well as $I/\sigma(I)$ were essentially those of random intensities. Focusing on the highest resolution shell, any level of JPEG-2000 compression generated measurable error. That said, we are interested in the measurable impact on the structural information that one can derive from the compressed data and comparing them with the uncompressed data. This will inform us as to the safe levels to use in future experiments.

Low levels of compression (50, 100, and 200) appear to alter data statistics only marginally. Most crystallographers would still find the data ‘good’ according to all the metrics included in Fig. 4 and Tables S4 and S5.

The ‘interesting’ area is with moderate JPEG-2000 requested compression ratios, 500, 1000 and 1500. J2K500

(~250 to 1 extra compression ratio) statistics are superficially sufficient to allow for structure solution. However, $I/\sigma(I)$ was significantly reduced, from 61 to 39. For J2K1500 (~750 to 1 extra compression ratio) the $I/\sigma(I)$ value is reduced drastically to 28.3 and reflection intensities and surrounding background levels were significantly impacted. Nevertheless, we attempted to solve the S_SAD lysozyme structure using all generated compressed data. Some of the over-compressed cases failed. The less-compressed successful cases are shown in bold in the supplementary data tables.

3.2.2. Hcompress compression

For Hcompress the amount of compression is specified as ‘scales’ rather than target compression ratios. Hcompress compression scales ranging from 4 to 64 were applied to the uncompressed data, corresponding to compression ratios ranging from 36 to 1344. Unlike JPEG-2000, which did not impact the overall completeness, Hcompress on the other hand results in gradual loss of completeness with increasing scale of compression. Further, R -factors are moderately impacted by the higher compression scales.

These two algorithms clearly display different effects on diffraction data that have been investigated at the various Hcompress reflection scales (see Section 4).

Table 1

A sampling of the successful processing of lysozyme compressed data sets, following the standard workflow: *XDS*, *pa_sg_res_cut* -S *P4₃2₁2*, *mtz2sca*, *aimless.mtz*, *aimless.sca*, and using *hkl2map* for *SHELXC/D/E* (merging and scaling / molecular replacement / density modification and autotracing).

Note that *pa_sg_res_cut* (https://github.com/dailekreitler-bnl/pa_sg_res) is a bash shell script wrapper for *aimless* and *pointless*, for the downstream processing of *fast_dp* output. This script allows for trimming integrated reflection files (from *XDS*) based on resolution and frame number, in addition to space group assignment. The number of heavy atom solutions from *SHELXD* (with CC weak greater than 15) is given, along with the number of residues traced by *SHELXE*, connectivity of map, dm.mtz FOM, number of residues autobuilt using *ARP/wARP*, and the final *ARP* model *R* value. See Tables S1, S2, and S3 in the supporting information for additional compression runs, many of which were successful and some of which failed.

Compression	Extra compression ratio	<i>SHELXD</i> solutions	<i>SHELXE</i> residues	<i>SHELXE</i> FOM / connectivity	DM FOM	<i>ARP</i> residues / correctness	<i>R</i>
Uncompressed	1	17	119	0.65 / 0.77	0.83	127 / 1.00	22.4
BIN2_SUM2_HCOMP04	17	2	106	0.65 / 0.76	0.83	127 / 1.00	23.5
BIN2_SUM2_HCOMP08	36	11	109	0.60 / 0.76	0.82	127 / 1.00	23.6
BIN2_SUM2_HCOMP16	112	15	103	0.63 / 0.76	0.82	127 / 1.00	23.1
BIN2_SUM2_HCOMP24	211	14	99	0.61 / 0.76	0.82	128 / 0.999	26.7
BIN2_SUM2_HCOMP32	323	10	112	0.65 / 0.76	0.81	127 / 1.00	23.7
BIN2_SUM2_HCOMP64	458	1	105	0.63 / 0.76	0.79	127 / 1.00	24.0

Learning from the compressions that were first tested for the HIV reverse transcriptase we limited compression algorithms and combinations to a selected set. Specifically, we did not test summing lysozyme by 5, 10, 20, and 40 frames or binning by 4 or 6 pixels. We focused on combinations of 2 × 2 binning and summing by 2 without and with various Hcompress scales and JPEG-2000 target compression ratios. Higher levels of summing in combination with 2 × 2 binning and Hcompress were tested to evaluate the impact of greater lossy compression for S_SAD phasing (see the labels to Fig. 4, panels 5 and 6).

3.2.3. Binning/summing/JPEG-2000

We applied JPEG-2000 compression to the ‘BIN2_SUM2’ frames with JPEG-2000 target compression ratios ranging from 25 to 200. Data statistics included in Fig. 4 and Tables S4 and S5 compared with the BIN2_SUM2 uncompressed data remained ‘unperturbed’ until we applied J2K140 or higher. Starting at target compression ratio 140, an increasing compression level produces *R*-factors and *I/σ(I)* statistics representing random diffraction with decreasing quality. It appears that the transition from level 130 to 140, for the lysozyme 7.5 keV data set for S_SAD phasing, corresponds to an inflection point. This transition will be investigated further in terms of Bragg reflections intensities and local background profile. These correspond to compression ratios of 1039 (level 130) and 1119 (level 140). Discussion of the effect of higher levels of compression from either JPEG-2000 or Hcompress is included below.

3.2.4. Binning/summing/Hcompress

As for the JPEG-2000 combined case, we tested six Hcompress scales, ranging from 4 to 64. Like the JPEG-2000 case, most compressed data processed with ‘unnoticeable’ damage when focusing on the parameters in Fig. 4 and Tables S4 and S5. We observe gradual noticeable decay of the high resolution statistics as well as the overall completeness and *I/σ(I)* but with no measurable impact on experimental phasing and subsequent automated model building. Compression ratios ranging from 34 to 915 were achieved.

Unlike for JPEG-2000, for Hcompress we also investigated the following combinations: BIN2_SUM5 and BIN2_SUM10 with Hcompress levels of 4, 16, and 64. These correspond to compression ratios ranging from 62 to 1336.

3.3. Lysozyme S_SAD phasing to test lossy compression

Starting with a S_SAD lysozyme structure, we investigated the effect of lossy compression of this data set on the solution of the S_SAD structure using a combination of four compression algorithms with compression ratios ranging from 50 to 2672. Two main algorithms were tested, JPEG-2000 and Hcompress, each with different impact. Not counting the native CBF data set, all 37 processed data sets were subsequently used for phasing experiments following the workflow derived from the uncompressed data.

As shown in Table 1 and Tables S1, S2, and S3 of the supporting information, the uncompressed data produced 17 correct *SHELXD* heavy atom searches. 37 lossy compressions were tested, with extra compression ratios ranging from 25 to 1336, and a subset of these are included in Table 1. A large number of the lossy compressions produced equally good heavy atom search solutions, with compression ratios up to almost 500:1. See Tables S1, S2, and S3 of the supporting information for additional compression runs, many of which were successful and some of which failed.

Two additional parameters, related to potential impact on phasing experiments, were included in Tables S4 and S5: the correlation of the intensities of two randomly selected half data sets between Friedel-mates (*CC*_{1/2} of anomalous data) and the mid-slope of the normal probability of anomalous difference, both provided by *AIMLESS*.

3.3.1. Uncompressed data S_SAD phasing and automated model building

The heavy atom search was performed searching for 14 S atoms, using 1000 trials unless otherwise specified. Phasing, density modification (DM) and chain auto-tracing were performed with 200 cycles of density modification with 40% solvent fraction and three cycles of tracing. DM was executed using the output from *SHELXE* for 200 cycles using 40%

solvent fraction and subsequently used for model building in *ARP/wARP* using default settings. The lysozyme data set was acquired at 7.5 keV to maximize anomalous contribution from the ten S atoms and the various Cl ions that are often observed when crystals are obtained from NaCl precipitant-based solution. In *SHELXD*, 17 out of the 1000 trials were successful sub-structure solutions with the best one saved and used in *SHELXE*. 119 out of 129 amino acids were traced in the density, which displayed a connectivity of 0.77 and a phase figure of merit (FOM) of 0.65. *ARP/wARP* placed and refined 127 amino acids including side chains with a crystallographic *R*-factor of 22.4. In other words, the uncompressed data resulted in straightforward experimental phasing and automated model building.

3.3.2. Compressed data S₂SAD phasing and model building

Phasing was attempted with all compressions tested and phasing results (pass or fail; pass defined as successful experimental phasing including heavy atom search, phasing and chain tracing) were used to perform further higher levels of compression and to explore additional combinations not available for the HIV reverse transcriptase. Here, it is important to distinguish two important steps relevant to successful phasing experiments: sub-structure solution followed by phasing and density modification. Of the 37 compressions tested (with ratios ranging from 8 to 1336), there were 27 successful runs that resulted in straightforward sub-structure solutions followed by essentially indistinguishable phasing results compared with the uncompressed data. These 27 compressions with ‘insignificant’ structural changes had ratios ranging from 8 to 1336.

One must emphasize that summing frames combined with binning pixels in those frames does significantly increase the intensities of all reflections with the most impact on partial strong reflections recorded over a few consecutive frames. When summing and binning were combined with Hcompress or JPEG-2000, higher compression ratios were achieved with minimum impact on phasing results. JPEG-2000 with target compression ratio greater than 1500, Hcompress with a scale of 64 as well as BIN2_SUM2 combined with JPEG-2000 J2Kxxx with xxx of 140 or greater did not yield successful sub-structure solution and phasing. Attempts to use up to 10000 *SHELXD* trials and good sub-structure solution failed.

A few compressions tested required up to 10000 *SHELXD* trials for a successful substructure solution needed for successful phasing and model building. These compressions were: HCOMP16 and HCOMP20 scales, BIN2_SUM5 with HCOMP4, as well as BIN2_SUM10 with and without HCOMP4.

The remainder of the compressions tested failed to yield sub-structure solutions. However, when using a set of known heavy atom positions from the previous *SHELXD* step, we were able to use *SHELXE* for phasing and phase improvement. In particular, as shown in Tables S1 and S2 of the supporting information, in the blue shaded rows, solution of the lysozyme data after lossy compression with JPEG-2000 at

target compression ratios of 500:1, J2K500, and 1000:1, J2K1000 (actual extra compression ratios of 251:1 and 498:1, respectively) needed the heavy atom positions from the solution to the JPEG-2000 200:1 target compression ratio, J2K200, compressed data (actual extra compression ratio 100:1), and solution of the lysozyme data after lossy compression with Hcompress at scales of 24, HCOMP24, and 32, HCOMP32 (actual extra compression ratios of 357:1 and 480:1, respectively) needed the heavy atom positions from the solution to the Hcompress scale 8, HCOMP08, compressed data (actual extra compression ratio 96:1).

Essentially, very high levels of compression impact diffraction peaks and background so that the data are unsuitable as observed during reduction and subsequent analysis for use in experimental phasing. The unsuitable cases were JPEG-2000 extra compression of 1000:1 with achieved extra compression ratio of 500:1, HCOMP64 achieving compression ratio of 336, and BIN2_SUM2_J2Kxxx with xxx greater than 140 corresponding to extra compression ratios greater than 225.

BIN2_SUM2 images generate stronger peaks than uncompressed data. Therefore BIN2_SUM2_J2Kxxx for higher values of xxx tends to impact counts on the shoulders of the reflections rather than the reflections themselves.

HCOMP64 significantly reduces the completeness of the data when applied to uncompressed data, but HCOMP64 when applied after BINx_SUMx does not significantly impact completeness.

That is, the achieved extra compression ratio is not, by itself, a good metric to determine whether or not phasing experiments can then be carried out. It is important to consider what compression algorithms are being used in what order and what they do to the reflections – see the discussion of reflections, below.

3.4. Compression results for HIV reverse transcriptase

The HIV reverse transcriptase data set presents a molecular replacement problem for phasing. Because we were looking for a ‘canary in a coal mine’ effect, we were unprepared for the resilience with which modern data reduction and structure solving software were able to handle highly compressed data. Our first benchmarking trial used total compression ratios only up to 200 (extra compression ratio 100), which proved insufficient to unambiguously probe the maximum limits of acceptable compression for two of our compression strategies (JPEG-2000 and HCOMPRESS). We then repeated our benchmarking to a compression ratio up to 500 (extra compression ratio 250). However, by combining multiple different compression strategies in sequence, we found that in some cases even 500 was insufficient to detect the onset of detectable degradation.

Hence, we repeated the entire exercise again with a maximum compression ratio for JPEG-2000 of 2000 (extra compression ratio 1000). Use of BIN2_SUM2 allowed us to reach higher extra compression ratios by a factor of as much as eight (see Figs. 5 and 6). Although we did not exhaustively test all possible combinations of compression strategies, we believe

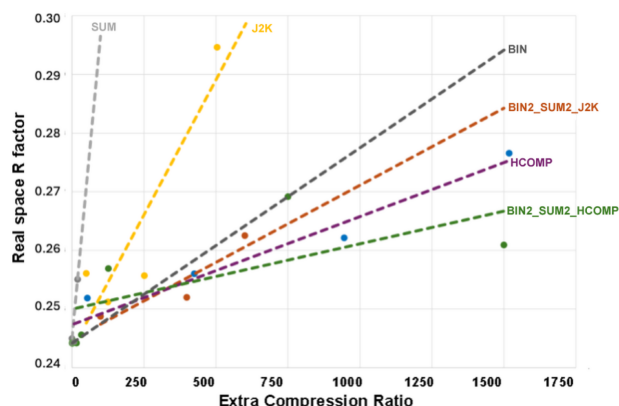


Figure 5

Benchmarking the effectiveness of compression strategies using real space R values for HIV reverse transcriptase. Real space R values were computed (i) using *overlapmap* and (ii) using custom software (the quotient in the former case is the total volume electron density, while in the latter case it is the variance in the electron density). The results of these two strategies were almost identical (see Tables S6 and S7); the average of the two measurements is plotted (Y axis) as a function of the compression ratio (X axis). It is evident from the image that increasing compression ratios are associated with decreasing map quality (increased real space R value), and the slope that governs this loss of quality is different for different compression strategies (the lower the slope, the better the compression strategy). In the case of the data examined here, the optimal strategy is to combine the effects of multiple compression strategies, i.e. binning two pixels in all three measured reciprocal space dimensions (x , y , and ϕ) followed by using the compression algorithm that was developed for astronomy purposes (this is labelled BIN2_SUM2_HCOMP above).

that the combination of twofold data binning followed by twofold frame binning followed by HCOMPRESS is the optimal strategy for lossy compression, which maximizes the total compression ratio with the least degradation of data quality.

An HIV Reverse Transcriptase Workflow is shown in the supporting information. The workflow was used to benchmark the effect of data compression on HIV RT data and on the resulting model (including a large 188 amino acid omit region).

In order to compute real space R values in a way that (i) permitted the comparisons of maps obtained using non-cognate atomic coordinates and (ii) permitted a custom definition of the real space R values, such that the denominator was the variance of the reference map, rather than the total integrated electron density, we apply the algorithm as shown in Fig. S1 of the supporting information.

In general, experimental phasing projects do not require an initial PDB model. On the other hand, ligand binding and molecular replacement projects do need an initial PDB model, and electron density maps are biased toward the model. For these experiments, one can afford higher compression ratios, compared with experimental phasing requiring more accurate reflection intensities.

Reducing model bias is a well established challenge in MX structures that are solved using molecular replacement (Ramachandran & Srinivasan, 1961; Wang, 1985). While it is tempting to presume that the signal-to-noise advantage would allow higher compression ratios of molecular replacement

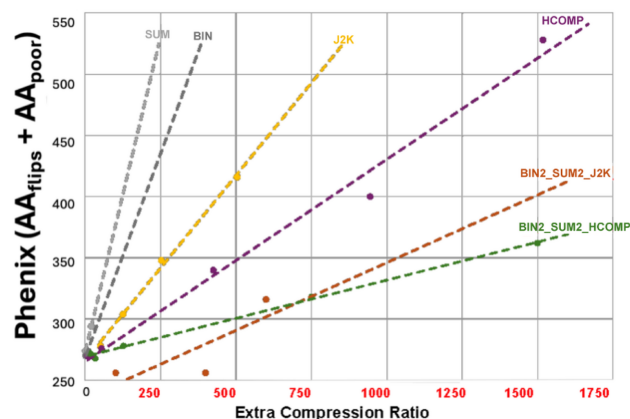


Figure 6

Benchmarking the effectiveness of compression strategies for HIV reverse transcriptase using binary metrics. The software package PHENIX (Adams *et al.*, 2010) was used to compare the electron density obtained from uncompressed data with the electron density obtained from data that were compressed using various strategies. The binary metrics (i) poor fit and (ii) conformational flip were added to indicate the total number of amino acids that were flagged as incorrect (hence the Y axis shows the number of amino acids that did not fit the observed electron density plus the number of amino acids that only fit with a different conformation). Just as in the case of Fig. 5, the slope of the decrease in quality varies between different compression strategies (and a smaller slope indicates a better compression strategy). Significantly, the best overall performance (labelled BIN2_SUM2_HCOMP) was the same when benchmarked using real space R values and when benchmarked using binary metrics.

experiments compared with experimental phasing experiments, we should be careful to ensure that model bias problems are not excessively increased at high compression ratios. To this end, we omitted a large domain from the model (188 residues between Thr^B₂₄₀ to GLN^B₄₂₈) before phasing each of the data sets that were compressed and then decompressed. We then were able to separately compute model correctness statistics (Table S8) and model-to-data fit statistics (Tables S9, S10, and S11 in the supporting information) for the entire protein and for the omitted region. We used these benchmarking results to quantify the impact of lossy data on model bias and to establish reasonable limits for prudent compression of primary data. Our results indicated that the electron density map obtained after molecular replacement remains reasonably correct in most regions when data are compressed by a factor of 500, even in the case of marginal data (although some errors may appear in large, omitted regions; see Fig. 7).

Table 2 shows statistics for two very conservative extra compression runs for HIV reverse transcriptase of up to 400:1. Tables S6 and S7 show more aggressive extra compression runs of as much as 1000:1 with satisfactory results.

4. Discussion

4.1. Hcompress versus JPEG-2000, impact of reflection profiles

The combined use of modest degrees of binning and summing strengthens the weak reflections, which helps to

Table 2

Sampling of successful HIV reverse transcriptase processing of compressed data, which worked for lossy compressions of more than 200:1.

Values in parentheses are for the highest resolution shell.

Compression	Extra compression ratio	R_m all	R_{pim} all	Mean $I/\sigma(I)$	Completeness	Total observations unique
Uncompressed	1	7.9 (144.8)	4.1 (72.9)	11.4 (1.0)	99.9 (100.0)	41 196 (4602)
BIN2_SUM2_J2K50	200	11.7 (273.1)	6.0 (137.3)	8.1 (0.6)	99.9 (100.0)	41 161 (4586)
BIN2_SUM2_J2K100	400	15.1 (319.4)	7.8 (161.0)	7.3 (0.5)	99.9 (100.0)	40 963 (4557)

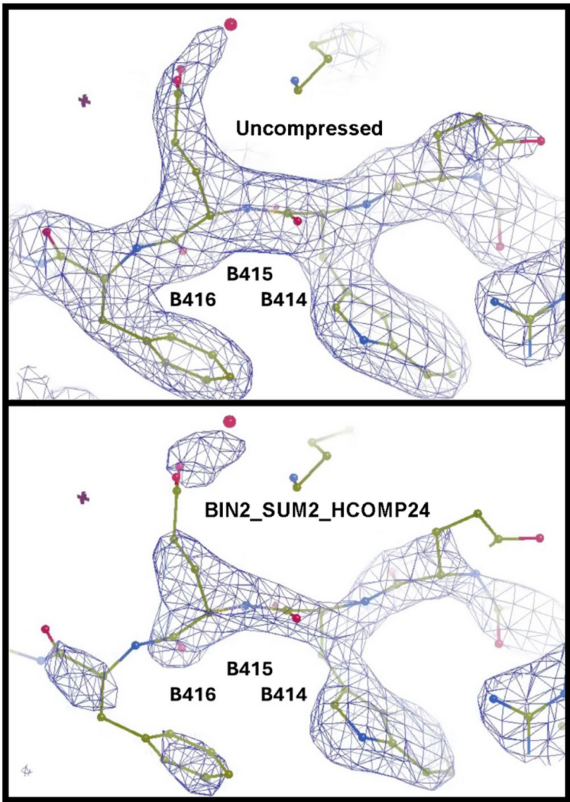


Figure 7

Electron density is shown for a region that contains some amino acids where the fit to the compressed data is good (right, B₄₁₄), moderate (middle, B₄₁₅), and poor (left, B₄₁₆). Compare the electron density in the omit region obtained using the uncompressed data (top image) and those obtained using the BIN2_SUM2_HCOMP24 data (bottom image), which has a compression ratio of 627. Electron density is shown for a region that contains some amino acids where the fit to the compressed data is good (right, B₄₁₄), moderate (middle, B₄₁₅), and poor (left, B₄₁₆). This region was selected specifically to demonstrate that (i) over-compression can initially manifest as difficult-to-detect locally poor fit between model and data, affecting only a few amino acids, and (ii) very aggressive compression ratios should be avoided in cases where good model-to-data fit is required for all amino acids in a structure.

protect them from the impact of the massive compressions provided by Hcompress or JPEG-2000. Figs. 8–11 show the impact of JPEG-2000 and Hcompress by themselves and in combination with binning and summing by two when working with the HIV reverse transcriptase data.

Fig. 8 shows the impact of JPEG-2000 compressions on weak reflections in HIV reverse transcriptase. J2K200 keeps the weak peaks findable but distorts the shoulders. The higher JPEG-2000 compressions do serious damage, losing the peak

entirely for J2K1000. Fig. 9 shows that Hcompress does better but also loses the peak entirely for high compressions. In Figs. 10 and 11 the weak peak has been maintained by modest binning and summing, thereby allowing useful extra compression ratios well over 250:1 to be achieved in many cases and over 500:1 in some cases.

The results we present reveal that, with care to protect weak reflections with modest binning and summing, and with care to avoid over-compression, in many cases massive lossy compression can be applied successfully while retaining the information necessary to find and integrate Bragg reflections and to solve structures with good statistics.

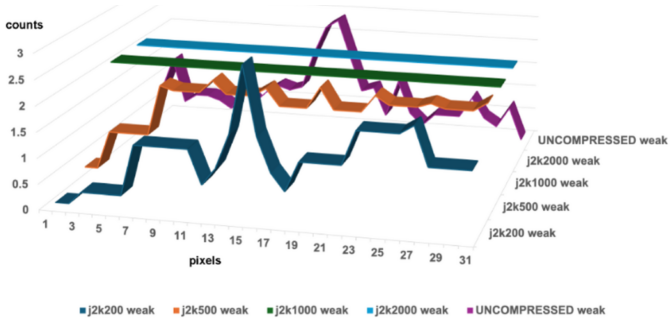


Figure 8

Impact of JPEG-2000 compressions at levels 200, 500, 1000, and 2000 on a typical weak reflection in HIV reverse transcriptase. Six rows through the reflection were averaged. The J2K200 compression preserves the peak but distorts the shoulders asymmetrically. The J2K500 compression severely distorts the profile, while J2K1000 and J2K2000 swallow the weak reflection completely.

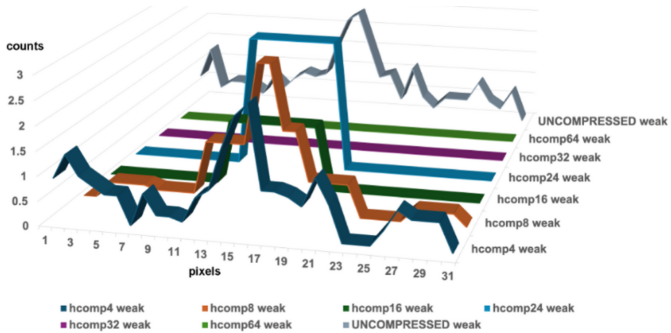


Figure 9

Impact of Hcompress compressions at scales 4, 8, 16, 24, 32, and 64 on a typical weak reflection in HIV reverse transcriptase. Six rows through the reflection were averaged. The hcomp4, 8 and 16 compressions preserve the peak, but the hcomp16 peak has been distorted. The hcomp24 compression actually boosted this weak reflection and hcomp32 and 64 swallowed it.

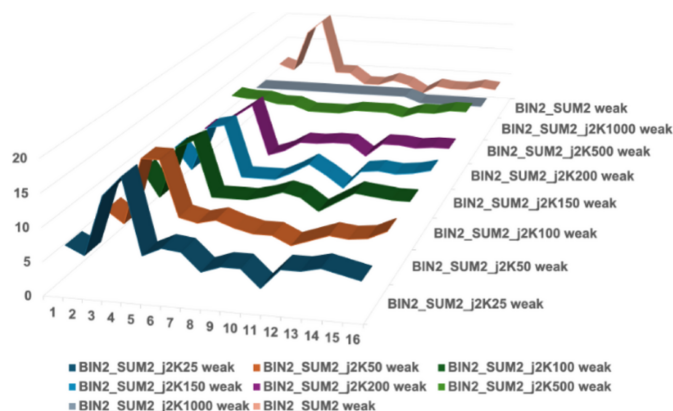


Figure 10

Impact of JPEG-2000 compressions at levels 25, 50, 100, 150, 200, 500, and 1000 on a typical weak reflection in HIV reverse transcriptase, first boosted and protected by binning and summing by factors of 2. Three rows through the boosted reflection were averaged. The BIN2_SUM2_J2K25, 50, 100, 150, 200, and 500 all preserve the basic shape of the weak reflection with some weakening that preserves the peak but distorts the shoulders asymmetrically. The J2K500 compression severely distorts the profile, while J2K1000 and J2K2000 swallow the weak reflection completely.

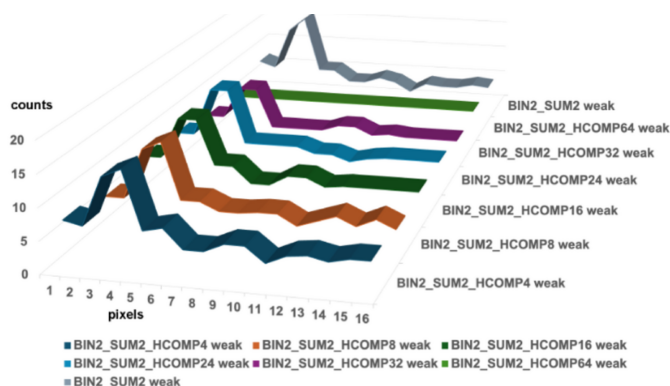


Figure 11

Impact of Hcompress compressions at scales 4, 8, 16, 24, 32, and 64 on a typical weak reflection in HIV reverse transcriptase first boosted and protected by binning and summing by factors of 2. Three rows through the boosted reflection were averaged. The BIN2_SUM2_HCOMP4, 8, 16, and 24 all preserve the basic shape of the weak reflection with no weakening or distortion. The HCOMP32 compression weakens the reflection slightly. The HCOMP64 compression swallows the weak reflection completely.

4.2. Suggestions for real time compression of MX data

We expect that new detectors with substantially higher framing rates combined with synchrotron-source upgrades and expected new sample-delivery methods will expectedly require the implementing of lossy compression. For many of the upcoming experiments, it will certainly be possible to engage lossy compression for real time processing depending on the type of experiment.

Using our initial results, we can extrapolate effects of binning ($4\times$ for 2×2) combined with either JPEG-2000 or

Hcompress for real time lossy compression. We would suggest that one consider BIN2_HCOMP32 or BIN2_JPEG125 with compression ratios of 750 or 500 for medium intensities and resolution, as for the HIV reverse transcriptase data set, and 150 or 250 for strong intensities and high resolution, as for the lysozyme data set. These compressions offer minimum loss of resolvable structural information regardless of the experiment. For greater compression ratios, we suggest BIN2_HCOMP64 offering compression ratios of 250 and 1200 for the strong and medium intensities.

4.3. Suggestions for long-term data archiving

For archiving of data, one can achieve significantly higher compression ratios by combining pixel binning ($4\times$ for 2×2), image summing ($2\times$, $4\times$, $5\times$) and subsequent image compression using either HCOMP64 or J2K125. One can achieve compression ratios of up to 500–1000 depending on the experiment. In other words, one can archive three years of a beamline operation (600 days) in the space originally required for 10 to 20 hours of data throughput. Highlighted in this way, it puts the compression discussion into a broader perspective of archiving data ‘forever’.

5. Availability of data and software

The lossless compression data for the lysozyme experiment have been deposited in Zenodo (<https://zenodo.org/records/12701763>) and the scripts in GitHub (https://github.com/nsls-ii-mx/lossy_compression_scripts). PDB depositions have been made for samples of the uncompressed and compressed structures of lysozyme. The entry from the uncompressed data is 9b7f. The entry for the data compressed with BIN2_CBF2_HCOMP24 is 9b7e.

6. Over-compression

While a surprisingly high degree of lossy compression that impacts the background and not the Bragg peaks is close to harmless for many macromolecular crystallographic experiments, there are risks in overdoing lossy compression. If the compression reaches into the pixels under the Bragg peaks or disturbs the relationship between the local background level and the Bragg peaks, peaks may be lost or ‘accidentally inflated’ in intensity. In the special case where diffuse scattering data is needed, requiring data from between the Bragg peaks, the best plan is both to keep some percentage of images (or possibly some percentage of complete data sets) faithfully compressed and to limit the level of lossy compression used on the remaining images to be within well tested limits derived from processing similar structures.

Every lossy compression technique has limits beyond which image statistics are severely degraded. It is prudent to avoid ‘over-compression’ and stay well within those limits. See the examples marked with yellow backgrounds in the supporting information.

7. Summary and conclusions

We employed three existing and well studied methods to accomplish dramatic compression where some information was lost, but much information was retained: binning of pixels, summing of frames, and JPEG-2000 from the movie industry or Hcompress from astronomy. For most MX images, one can almost always achieve at least 2:1 lossless compression, as provided by the detector control unit. The combination of 2×2 binning, summing by 2, and Hcompress was particularly effective in achieving useful extra lossy compressions of between 250:1 and 500:1 when the primary focus of the experiment is to process Bragg-reflection data. As a result, one can reliably reduce the storage by a factor of better than 300:1. This will allow us to be sure to save enough data to be able to detect cases for which non-Bragg background data will need to be processed, so that very effective compression ratios will be achievable for the transition to much faster detectors. In this discussion we are not addressing throughput but are focusing on achievable compression ratios and the effect of the compressions used on the quality of the structures produced.

These results encourage us to study lossy compression further using crystallographic knowledge to subtract the whole background. In many cases we would compress only pixels directly associated with Bragg peaks. That will allow one to preserve important experimental data with overall compression ratios of between 500:1 and 10000:1 in many high-data-rate macromolecular experiments.

8. Recapitulation discussion

Lossy compression techniques can significantly reduce data size by focusing on retaining essential properties while sacrificing less-critical ones. Key questions to address are: (i) which data properties should be preserved, (ii) which compression methods best maintain these properties, and (iii) how to collect data to achieve high compression ratios. Future work should explore these questions, including techniques such as developing better diffuse scattering techniques, matching compression strategies to specific data types, and optimizing data collection methods. For example, the results of our experiments suggest considering use of larger detector pixels and larger rotations for each image. This discussion aims to ignite a dialogue on lossy compression for MX data, recognizing that it may affect different community members unevenly and valuing diverse perspectives.

Acknowledgements

We gratefully acknowledge thoughtful comments and suggestions by Robert M. Sweet and Frances C. Bernstein. The entire community owes a debt to James Holton for his creative, useful, and constructive comments and suggestions on the subject of lossy compression for macromolecular crystallography since 2011. He helped us all to break free of the past and look dispassionately at what modern realities require. We thank Dr Karen Anderson for generously

allowing us to use her HIV reverse transcriptase data to benchmark our compression algorithms, and for her comments regarding the ways that small changes in data quality could affect the interpretability of data in the vicinity of key areas of the protein that are biologically significant.

Funding information

This work utilized the AMX beamline from the Center for BioMolecular Structure (CBMS) which is primarily supported by the National Institutes of Health, National Institute of General Medical Sciences (NIGMS) through a Center Core P30 Grant (P30GM133893), and by the DOE Office of Biological and Environmental Research (KP1607011). As part of NSLS-II, a national user facility at Brookhaven National Laboratory, work performed at the CBMS is supported in part by the US Department of Energy, Office of Science, Office of Basic Energy Sciences Program under contract number DE-SC0012704.

References

- Abrahams, J. P. (1993). *Jnt CCP4 ESF-EACBM Newsl. Protein Crystallogr.* **28**, 3–4.
- Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W., Echols, N., Headd, J. J., Hung, L., Kapral, G. J., Grosse-Kunstleve, R. W., McCoy, A. J., Moriarty, N. W., Oeffner, R., Read, R. J., Richardson, D. C., Richardson, J. S., Terwilliger, T. C. & Zwart, P. H. (2010). *Acta Cryst.* **D66**, 213–221.
- Berman, H., Henrick, K. & Nakamura, H. (2003). *Nat. Struct. Mol. Biol.* **10**, 980.
- Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
- Bernstein, H. J. (2010). *imgCIF, HDF5, NeXus: Issues in Integration of Images from Multiple Sources at HDF5 as Hyperspectral Data Analysis Format Workshop*, 11–13 January 2010, ESRF, Grenoble, France.
- Bernstein, H. J., Andrews, L. C., Diaz, J. A. Jr, Jakoncic, J., Nguyen, T., Sauter, N. K., Soares, A. S., Wei, J. Y., Wlodek, M. R. & Xerri, M. A. (2020). *Struct. Dyn.* **7**, 014302.
- Bernstein, H. J., Ellis, P. J., Hammersley, A. P., Howard, A. J., Pflugrath, J. W., Sweet, R. M. & Westbrook, J. D. (1999) *Acta Cryst.* **A55**, 235.
- Bernstein, H. J. & Jakoncic, J. (2024). *J. Synchrotron Rad.* **31**, 647–654.
- Brönnimann, C. (2005). *Acta Cryst.* **A61**, C21.
- Brönnimann, C., Eikenberry, E. F., Horisberger, R., Hülsen, G., Schmitt, B., Schulze-Bries, C. & Tomizaki, T. (2003). *Nucl. Instrum. Methods Phys. Res. A*, **510**, 24–28.
- Chan, A. H., Duong, V. N., Ippolito, J. A., Jorgensen, W. L. & Anderson, K. S. (2020). *6X4C*, <https://doi.org/10.2210/pdb6x4c/pdb>.
- Collet, Y. (2011). *lz4*, <https://github.com/lz4/lz4>.
- Collet, Y. & Kucherawy, M. (2021). *RFC 8878 Zstandard Compression and the 'application/zstd' Media Type*, <https://www.rfc-editor.org/rfc/rfc8878.pdf>.
- Cowtan, K. (2010). *Acta Cryst.* **D66**, 470–478.
- Data, R. B. (1947). *Proc. Leeds Philos. Soc.* **5**, 1–13.
- Donath, T., Rissi, M. & Billich, H. (2013). *Synchrotron Radiat. News*, **26**(5), 34–35.
- Duong, V. N., Ippolito, J. A., Chan, A. H., Lee, W. G., Spasov, K. A., Jorgensen, W. L. & Anderson, K. S. (2020). *Protein Sci.* **29**, 1902–1910.

- Ellis, P. J. & Bernstein, H. J. (2006). *International Tables for Crystallography* Vol. G, edited by S. R. Hall & B. McMahon, pp. 544–556. Heidelberg: Springer.
- Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* **D60**, 2126–2132.
- Evans, P. (2014). *J. Crystallogr. Soc. Japan*, **56**, s24.
- Evans, P. R. & Murshudov, G. N. (2013). *Acta Cryst.* **D69**, 1204–1214.
- Ferrer, J.-L., Roth, M. & Antoniadis, A. (1998). *Acta Cryst.* **D54**, 184–199.
- Förster, A., Brandstetter, S., Müller, M. & Schulze-Briesse, C. (2016). *EIGER detectors in biological crystallography*, White Paper, https://media.dectris.com/White_Paper_EIGER_May2016_hPpzdJD.pdf.
- Fritze, K., Lange, M., Möstl, G., Oleak, H. & Richter, G. M. (1977). *Astron. Nachr.* **298**, 189–196.
- Galchenkova, M., Tolstikova, A., Klopprogge, B., Sprenger, J., Oberthuer, D., Brehm, W., White, T. A., Barty, A., Chapman, H. N. & Yefanov, O. (2024). *IUCrJ*, **11**, 190–201.
- Hiraki, T. N., Abe, T., Yamaga, M., Sugimoto, T., Ozaki, K., Honjo, Y., Joti, Y. & Hatsui, T. (2021). *Acta Cryst.* **A77**, C531.
- Hollander, K., Chan, A. H., Frey, K. M., Hunker, O., Ippolito, J. A., Spasov, K. A., Yeh, Y. J., Jorgensen, W. L., Ho, Y. C. & Anderson, K. S. (2023). *Protein Sci.* **32**, e4814.
- ISPGGroup (2023). *openjpeg*, <https://github.com/uclouvain/openjpeg>.
- Kabsch, W. (2010). *Acta Cryst.* **D66**, 125–132.
- Kroon-Batenburg, L. M. J. & Helliwell, J. R. (2014). *Acta Cryst.* **D70**, 2502–2509.
- Langer, G., Cohen, S. X., Lamzin, V. S. & Perrakis, A. (2008). *Nat. Protoc.* **3**, 1171–1179.
- Lelewer, D. A. & Hirschberg, D. S. (1987). *ACM Comput. Surv.* **19**, 261–296.
- Masui, K., Amiri, M., Connor, L., Deng, M., Fandino, M., Höfer, C., Halpern, M., Hanna, D., Hincks, A. D., Hinshaw, G., Parra, J. M., Newburgh, L. B., Shaw, J. R. & Vanderlinde, K. (2015). *arXiv:1503.00638*.
- Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* **D53**, 240–255.
- Pape, T. & Schneider, T. R. (2004). *J. Appl. Cryst.* **37**, 843–844.
- Pence, W. D., Chiappetti, L., Page, C. G., Shaw, R. A. & Stobie, E. (2010). *Astron. Astrophys.* **524**, A42.
- Poynton, C. A. (1992). *Proc. SPIE*, **1659**, 152–158.
- Prucha, G. R., Henry, S., Hollander, K., Carter, Z. J., Spasov, K. A., Jorgensen, W. L. & Anderson, K. S. (2023). *Eur. J. Med. Chem.* **262**, 115894.
- Rahman, M. A. & Hamada, M. (2019). *Symmetry*, **11**, 1274.
- Ramachandran, G. N. & Srinivasan, R. (1961). *Nature*, **190**, 159–161.
- Santa-Cruz, D., Grosbois, R. & Ebrahimi, T. (2002). *Signal Process. Image Commun.* **17**, 113–130.
- Schneider, D. K., Soares, A. S., Lazo, E. O., Kreidler, D. F., Qian, K., Fuchs, M. R., Bhogadi, D. K., Antonelli, S., Myers, S. S., Martins, B. S., Skinner, J. M., Aishima, J., Bernstein, H. J., Langdon, T., Lara, J., Petkus, R., Cowan, M., Flaks, L., Smith, T., Shea-McCarthy, G., Idir, M., Huang, L., Chubar, O., Sweet, R. M., Berman, L. E., McSweeney, S. & Jakoncic, J. (2022). *J. Synchrotron Rad.* **29**, 1480–1494.
- Shannon, C. E. (1948). *Bell Syst. Tech. J.* **27**, 379–423.
- Sheldrick, G. M. (2008). *Acta Cryst.* **A64**, 112–122.
- Underwood, R., Yoon, C., Gok, A., Di, S. & Cappello, F. (2023). *Synchrotron Radiat. News*, **36**(4), 17–22.
- Vagin, A. A., Steiner, R. A., Lebedev, A. A., Potterton, L., McNicholas, S., Long, F. & Murshudov, G. N. (2004). *Acta Cryst.* **D60**, 2184–2195.
- Wang, B. C. (1985). *Methods Enzymol.* **115**, 90–112.
- Welch, T. A. (1984). *Computer*, **17**, 8–19.
- Welch, T. A. & Sperry Corp (1985). *High speed data compression and decompression apparatus and method*. US Patent 4558302.
- White, R. L. (2019). *Hcompress Image Compression Software*, <https://www.stsci.edu/software/hcompress.html>
- White, R. L., Postman, M. & Lattanzi, M. G. (1992). *Compression of the guide star digitised Schmidt plates*. In *Digitised Optical Sky Surveys: Proceedings of the Conference on 'Digitised Optical Sky Surveys'*, 18–21 June 1991, Edinburgh, Scotland, pp. 167–175. Springer Netherlands.
- White House (2022). *OSTP Issues Guidance to Make Federally Funded Research Freely Available without Delay*, <https://www.whitehouse.gov/ostp/news-updates/2022/08/25/ostp-issues-guidance-to-make-federally-funded-research-freely-available-without-delay/>.
- Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., Gonzalez-Beltran, A., Gray, A. J. G., Groth, P., Goble, C., Grethe, J. S., Heringa, J., 't Hoen, P. A. C., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S. J., Martone, M. E., Mons, A., Packer, A. L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M. A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J. & Mons, B. (2016). *Sci. Data*, **3**, 160018.
- Winter, G. & McAuley, K. E. (2011). *Methods*, **55**, 81–93.
- Ziv, J. & Lempel, A. (1977). *IEEE Trans. Inf. Theory*, **23**, 337–343.