

XRAIS: Physics-informed artificial intelligence for monitoring crystallography experiments

Derek Mendez,^a James M. Holton,^{a,b,c} Artem Y. Lyubimov,^a Sabine Hollatz,^a Irimpan I. Mathews,^a Alexander Cichosz,^d Vardan Martisoyan,^d Teo Zeng,^d Ryan Stofer,^d Robin Liu,^d Jinhui Song,^a Scott McPhillips,^a Mike Soltis,^a Aina E. Cohen^a

^a Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, Menlo Park, CA 94025, USA

^b Molecular Biophysics and Integrated Bioimaging Division, Lawrence Berkeley National Laboratory, Berkeley, CA 94720, USA

^c Department of Biochemistry and Biophysics, UC San Francisco, San Francisco, CA 94158, USA

^d Department of Statistics and Applied Probability, UC Santa Barbara, Santa Barbara CA 93106, USA

Synopsis Artificial intelligence was used to characterize the diffraction in images from serial and rotation crystallography experiments. Forward simulations were used to train models to infer B-factors, resolutions, and the presence of crystal splitting from single diffraction images.

Abstract The rising X-ray diffraction data acquisition rates make evaluation of collected images a challenge for experimenters. The use of artificial intelligence to evaluate diffraction images can help, but assembling large and precisely designed training datasets is difficult. To address this, we developed a suite of tools called XRAIS (X-ray Artificial Intelligence from Simulation) for simulating diffraction data, training models on those data, and using those models to monitor crystallography experiments. Here we demonstrate the capability of XRAIS-derived models to correctly interpret crystal resolution and the presence of multi-lattice diffraction across a compilation of diffraction images from 25 synchrotron experiments and 2 X-ray free electron laser experiments. XRAIS models can now provide real-time feedback on resolution and multi-lattice scattering to users of the macromolecular crystallography instruments at the Stanford Synchrotron Radiation Lightsource. This work highlights the utility of physics-based simulation in the context of artificial intelligence and lays the groundwork for development of additional models to better automate diffraction collection and analysis.

Keywords: **artificial intelligence, serial crystallography, rotation crystallography**

1. Introduction

Crystallography data rates are on the rise at synchrotrons (SRs) and X-ray free electron lasers (XFELs) alike. At SRs, high brilliance undulator beamlines coupled with advances in robotics and detector technologies have accelerated the pace of experiments, requiring faster algorithms to provide feedback on experimental outcomes. For example, at the microfocus beamline 12-1 of the Stanford Synchrotron Radiation Lightsource (SSRL), datasets may be collected with crystal rotation speeds up to 90° per second and frame rates exceeding 100 Hz (Cohen 2021). Moreover, at XFEL facilities, pulses are inherently short (on the order of femtoseconds), and each contain on the order of 10^{12} photons. During serial crystallography experiments at XFELs, crystals are exposed in rapid succession to produce hundreds to millions of snapshot diffraction images devoid of radiation induced artifacts. Hard X-ray pulses are produced at 120 Hz at the Linac Coherent Light Source (LCLS), the first XFEL, and similar at SACLA (Nango et al 2019), PAL (Park et al 2016) and SwissFEL (Milne et al 2017). At the European XFEL, due to advances in radio frequency accelerator technology (Singer et al 2015), hard X-ray pulses can be produced at 27 KHz (Weidorn et al 2018), and with similar technology, the up-and-coming LCLS-II facility is aiming to exceed this (Antipov et al 2018, Raubenheimer 2018). While high resolution diffraction images cannot currently be collected this fast, at XFELs the AGIPD (Allahgholi et al 2019) and JUNGFRAU 16M (Leonarski et al 2018) can record megapixel diffraction images at 3.5KHz and 1.1 KHz, respectively, and at SRs the Dectris EIGER 16M can record at 3 KHz (Casanas et al 2016).

Prompted by these advances, crystallography projects using SR and/or XFEL sources are transitioning from the determination of a single static structure to the application of time-resolved serial methods. In time-resolved crystallography, experimenters obtain multiple structures that illuminate details of the atomic positions and motions during biological processes (Schulz et al 2022, de Wijn et al 2022, Brändén and Neutze 2021, Pearson and Mehrabi 2020, Nango et al 2019, Pandey et al 2019, Šrajer and Schmidt 2017, Tanaka et al 2017, Schmidt 2015). During these experiments, crystals are transported into the X-ray beam position at unprecedented speed using serial sample delivery methods that apply fast translations of high-density crystal holders or liquid-crystal injectors (Deponte et al 2008, Fuller et al 2017, Martiel et al 2019, Baxter et al 2016, Lieske et al 2019). However, full utilization of the exciting advantages afforded by these technological leaps will require additional tools for real-time diffraction image analysis. Furthermore, without feedback, expensive and limited-quantity samples may be used up during experimental optimization or

lost if problems occur (e.g., with sample positioning or X-ray beam alignment). The implication here is the urgent need for someone (or something) to analyze incoming images and quickly make judgment calls about diffraction quality.

Work towards this goal has already progressed. For example, in (Ke et al 2018) the authors trained a convolutional neural network to screen images for any signs of diffraction, for hypothetical use in a program that could quickly reject so-called “misses” (i.e., images without diffraction, which comprise significant percentages of data collected using high-flow rate injector methods). These “misses” could then be excluded from processing and/or recording to disk to free up computing resources. More recently, in (Rahmani et al 2023), the authors made use of various dimensionality reduction algorithms to convert diffraction data into a set of features suitable for training a machine learning classifier to automatically detect whether experimental images contained diffraction.

The above methods are applicable to scenarios where a large fraction of images contain misses (e.g., liquid injection experiments), however there are also many high framerate experiments where every image contains diffraction, for example fixed target serial crystallography (Lieske et al 2019, Baxter et al 2016, Cohen et al 2014) and high-speed rotational crystallography (Cohen et al 2021). We propose here to move beyond the binary detection of diffraction, and to use artificial intelligence (AI) based computer algorithms to describe the observed scattering. Two common questions crystallographers want to know during experiments are: 1) “What is the diffraction resolution?”, and 2) “Does my image contain multi-lattice scattering?”. Crystallographers are good at answering these questions using visual inspection, but this practice is inefficient and infeasible at high data rates. Conventional crystallographic algorithms can answer the first question but are sensitive to input parameters and image artifacts. For example, the resolution estimation program implemented in the DIALS software suite (Winter et al 2018) is sensitive to image artifacts from ice diffraction and so-called “hot pixels”. Further, multi-lattice scattering, a common hindrance to data processing software, is exceedingly difficult to detect using conventional methods, with detection usually requiring determination of crystal orientations and unit cell dimensions (Gildea et al 2014, Schmidt 2014).

For the presented work, forward simulation software was used to create vast and diverse training datasets of X-ray diffraction images. Crucially, these images were automatically labeled according to the underlying physics. Subsequently, modern AI-based optimization techniques were used to train a regression model to predict resolution, and a classification

model to predict multiple-lattice scattering. Both models accepted a two-dimensional diffraction pattern as input, after applying a simple downsampling filter. The trained XRAIS models are now incorporated at SSRL beamlines in an automated framework that analyzes every collected image and displays the results to users in the Blu-ice/DCSS GUI (McPhillips et al 2002). Before XRAIS, this framework relied on computational methods implemented in DIALS using default parameters, but in many instances default parameters were inadequate. XRAIS models on the other hand are parameterless, and better suited for automated applications. During rotation data collection, XRAIS inferences can also be used to monitor for radiation damage, crystal mis-centering, and non-isomorphous diffraction. During serial experiments at SSRL BL12-1, the results can be used to optimize experimental parameters such as injector flow rate, X-ray attenuation, and/or beam size. XRAIS can correctly interpret images from a wide variety of sources and detectors. We intend to expand XRAIS to predict even more parameters of interest, and to produce a stable framework for general use at X-ray crystallography facilities worldwide.

2. Methods

2.1. Simulating training data

The goal of this work is to accurately predict parameters that describe diffraction quality using models derived from simulated data. To generate training data from which to build prediction models, we used the program *nanoBragg* (Holton et al 2014, Lyubimov et al 2016, Sauter et al 2020), which simulates crystal diffraction according to the kinematic theory of diffraction (James 1962). *nanoBragg* incorporates user-defined background scattering and adds noise by sampling Poissonian and Gaussian distributions describing photon counting and electronic noise, respectively. The use of simulated images facilitates the creation of large training datasets that would be impractical to accurately label by hand. Furthermore, it becomes possible to create training datasets that vary or isolate any combination of properties. For all the simulations reported here, a variety of parameters were randomly sampled, including detector distances, detector types, beam-stop sizes, bad-pixel masks, hot-pixel masks, proteins, space groups, unit cells, crystal volumes, mosaic spreads, and background scatter. These are summarized in Appendix A1. For each simulated image, only one quadrant was used for training (the upper left) and stored as a maxpool-downsampled array of 512 x 512 pixels.

2.1.1. Resolution training data

Resolution is perhaps the most important quality metric in any structural biology experiment because it defines the clarity of the structural image. Formally, resolution is the minimum separation distance between two features required for those two features to be identified as distinct from one another, e.g., at 1 Å resolution, individual atoms can be clearly resolved, whereas at poorer resolutions (2 - 3 Å), amino acid side chains are resolvable, but individual atomic positions must be inferred from prior knowledge and are less reliable. In practice, X-ray crystallographers determine the resolution cutoff at the point at which the merged diffraction data become uninterpretable. Criteria for inferring resolution have evolved over the decades. Oftentimes the recent and widely accepted CC1/2 metric defined by Karplus and Diederichs (2012) is used to set the resolution. The related signal to noise ratio of the structure factor intensities was used to define a resolution in this work for comparison with XRAIS inferences.

The “*resolution of a diffraction pattern*” is also a concept commonly used when discussing X-ray diffraction experiments themselves and is defined by the widest angle from the incident beam at which Bragg peaks can be observed. Observability of the Bragg peaks is in turn related to the rate at which the diffraction decays on the image, parameterized by a quantity called the *B-factor* (Bragg 1914). Higher B-factors indicate disorder in the crystal due to uncertainties in atomic positions, which ultimately affects the resolution of a dataset, causing diffraction to fall off more rapidly with resolution and obscuring reflections at wider scattering angles. B-factors and resolutions are included with structures deposited in the Protein Data Bank (PDB) (Berman et al 2000), which makes them amenable to data mining. Thus, an analysis of B-factors and resolutions revealed a simple non-linear relationship first described in (Holton 2009) and shown here in Figure 1.

With this relationship as an underlying assumption, a resolution prediction training dataset was created by simulating images with varying B-factors. Figure 2 shows a randomly selected assortment of these simulated images and their corresponding resolutions. Note that the resolution cutoff doesn’t always align with the point at which the diffraction becomes invisible in the image. Instead, resolution is defined by the rate of diffraction intensity decay as expressed by the B-factor. But it is complicated by the varying degrees of background in each image: a high-resolution image can also have a high background that makes it appear to be a low-resolution image (Figure 2F), adding uncertainty to our training data labels. Further, specific to synchrotron experiments, the dose received by a crystal also influences the B-factor (Holton 2009, Kmetko et al 2006), and ultimately the resolution. A strategy to account

for these additional factors is described in (Holton and Frankel 2010), but for the main results presented here, we rely on the generality of the relationship between B-factor and resolution shown in Figure 1 and note that the B-factor is the dominant term affecting the damage-limited intensity from a protein crystal, appearing as a Gaussian expression in (Holton and Frankel 2010, equation 18 therein). Resolution training data were simulated on a combination of PILATUS 6M and EIGER 16M camera models with variable detector distances on the range 200 - 300 mm. All simulations assumed a fixed photon energy of 0.9795 Å. See Appendix A1 for further details.

2.1.2. Multi-lattice training data

Multi-lattice scattering occurs when multiple crystal domains are exposed simultaneously, either because the diffracting volume contains a crack or a major dislocation, or if several crystals are caught in the beam. This effect undermines diffraction data processing software algorithms, which, for the most part, assume that diffraction comes from a single lattice. To simulate training data for multiple lattice scattering, a random number of lattices (1, 2 or 3) was “placed” before the simulated X-ray beam at randomized orientations. For this training, rotational mosaic spread was kept small (< 0.01°), and lattice orientations were drawn from a Gaussian distribution with a randomly chosen variance of (0.1°, 1°, or 10°), and a mean of 0° (about the nominal crystal orientation). In this way, it was theoretically possible for Bragg peaks from different lattices to closely overlap in a single image, thus simulating diffraction from a cracked crystal. Figure 3 shows a randomly selected assortment of multi-lattice training data and illustrates how image features vary with the number of lattices. Training data for this model used a Rayonix 345 (Rayonix L.L.C.) detector format matching the geometry from an LCLS experiment (Artz et al 2020), however it was found that the model generalized well to other datasets (as described in Section 3). The training dataset was made up of 50% single lattice images, 25% two-lattice images, and 25% three-lattice images.

2.2. Image conditioning

All images (both simulated and experimental) were downsampled and normalized before model evaluation, as the raw data formats considered for this study (Dectris PILATUS 6M, Rayonix 345, JUNGFRAU 16M, Dectris EIGER 16M) are large. To downsample an image by a factor of N (N=2 for PILATUS 6M, N=4 for EIGER 16M, JUNGFRAU 16M, and Rayonix), raw pixels were grouped into N x N blocks, and the value of each “block pixel” was set as the maximum value of the N² raw pixels inside of it. The downsampled “block

“pixel” values were then replaced by their square root and cast as integers. This data conditioning process is shown in detail in Figure 4 for a region of a PILATUS 6M image containing a Bragg reflection. After downsampling, images were divided into 4 quadrants of size 512 x 512 pixels, each of which could be passed to our AI trained models to produce independent estimates for predictors. For training, only one quadrant was used from each simulated crystal diffraction image. For experimental prediction, multiple quadrants were used as discussed below, but only one quadrant was required, and this case should be considered depending on available compute resources.

2.3. Model fitting

PyTorch (Paszke et al 2017) was used to fit regression (resolution prediction) and classification (multi-lattice detection) models using our training datasets. In general terms, PyTorch was tasked with reducing the error (“loss”) between the ground truth labels and the ones derived from the current model. For resolution prediction training, the loss function was the Mean Squared Error between labels and predictions in inverse units, i.e., inverse resolution was predicted by the model and compared to inverse resolution labels (e.g., an image simulated with B-factor corresponding to 2 Å resolution was labeled by 0.5 \AA^{-1}). For multi-lattice detection training, the Binary Cross Entropy loss function was used. Training labels were set to 0 or 1 (single- or multi- lattice scattering), and model predictions were mapped to a probability using a Sigmoid function, and then rounded to 0 or 1 before computing the loss.

2.3.1. Model Architecture

Currently, XRAIS uses a ResNet (He et al 2015) architecture with a modified input/output stage for predicting resolution and detecting multiple lattices. ResNet is a state-of-the-art deep convolutional neural network architecture, which accepts RGB images as inputs. For each image, it outputs 1000 numbers (features) intended for use in a multi-class classification model (with up to 1000 possible outcomes). To use ResNet with diffraction images, its input layer (a convolutional layer) was modified to accept single-channel (greyscale) images. Secondly, as originally done in (LeCun et al 1998), two fully connected (FC) layers were chained together at the output stage to convert the 1000 numbers into a single number suitable for prediction. The first FC layer mapped 1000 numbers to 100 numbers using 10^5 parameters, while the second FC layer mapped 100 numbers to 1 number (using 10^2 parameters). Also, following (LeCun et al 1998), a rectified linear unit activation function

was used between the first and second FC layer (see Figure 5), adding nonlinearity to the FC models. Figure 5A shows the baseline architecture used for both resolution and multi-lattice prediction models. Each model has unique aspects related to the desired predictor. For resolution, an additional input vector of basic diffraction geometry quantities (detector distance, pixel size, wavelength) was used to convert the output of the base model to an inverse-resolution quantity (Figure 5B). Modeling inverse resolution prevented scenarios where 0-division could occur during model training. For multi-lattice detection, a Sigmoid function was used to convert the output to the range 0-1, typical for binary classification (Figure 5C).

2.3.2. Model training

For the resolution prediction model, we trained on a dataset comprising 200,000 PILATUS 6M and 125,000 EIGER 16M images, each labeled with a unique resolution according to its B-factor, and with a randomized sample-to-detector distance. After each epoch (a pass through the entire training set, computing the loss function and its gradient for every training example), the model was validated on 10% of the simulated images (set-aside for testing and not included in training). The resolution inference training loss curve is shown in Figure 6A for both train and test sets. Training was carried out on 16 Perlmutter GPU nodes at NERSC, utilized 64 A100 GPUs, and ran at a speed of 0.7 minutes per epoch. For the multi-lattice detection model, training was done using 117,000 simulated diffraction images, each labeled by a Boolean indicating the presence of multiple lattices, and at each epoch the model was validated on 13,000 simulated images (Figure 6B). Training was carried out on 10 Cori GPU nodes at NERSC, utilizing 80 V100 GPUs, and ran at a speed of 1.6 minutes per epoch. Multi-node training at NERSC was done using Pytorch's Distributed Data Parallel protocol. Training on a single GPU machine was also tested, and, using a single V100 GPU, training a 43,000-image training took 11.5 minutes per epoch. When training on a single GPU, fewer epochs were required to reach convergence, and the full utility of Distributed Data Parallel is still being investigated. Table 2 summarizes the hyper-parameters and architectures used for both models.

3. Results

3.1. Resolution prediction in JUNGFRAU16M SwissFEL data

The resolution model was tested on a serial JUNGFRAU 16M dataset collected at SwissFEL light source. There, CYP121 crystals (Fielding et al 2017) were introduced to the SwissFEL

SASE (not pink) beam using a tape-drive setup (Fuller et al 2017) operated at ambient temperature and pressure. Each JUNGFRAU diffraction image was written to disk as a three-dimensional array (32 x 1024 x 512 pixels) however our resolution prediction model expected 512 x 512 quadrant images, oriented with the beam center aligned with the first pixel in memory (e.g., as in Figure 2, 3). To accommodate the model, each JUNGFRAU image was cast as a two-dimensional array of size 4434 x 4218, and afterwards, the data were downsampled into 512x512 quadrants (Section 2.2). A resulting JUNGFRAU quadrant is shown in Figure 7A. Figure 7 describes the results from XRAIS inferring resolution for the entire dataset of 9,592 images. The predicted resolutions were in the range of 1.3 – 5.7 Å (Figure 7D), and the resolution obtained from *cctbx.xfel.merge* after processing all 9,592 was 1.6 Å. It's noteworthy that the resolution model used here was trained on PILATUS 6M and EIGER 16M geometries, but it worked out-of-the-box for these JUNGFRAU 16M data.

3.2. Resolution prediction for SSRL data:

XRAIS resolution inference for 25 rotation datasets obtained at SSRL/SMB was also performed. Table 3 describes these datasets. Each dataset was labeled by an overall resolution, determined from the output of *Aimless* (Evans and Murshudov 2013) as the point (resolution) where the signal-to-noise ratio of the structure factor intensity dipped below 1.5. Figure 8 shows the XRAIS resolution versus image number for each of these datasets. For each diffraction image, we predicted 4 resolutions (1 per quadrant), and took either the minimum, or the mean resolution across the quadrants as the effective resolution (what's displayed in the plot as red or blue markers). Also shown in Figure 8 is the per-image resolution estimated by *DIALS* (Winter et al 2018). In most cases XRAIS inference worked qualitatively well, and it was shown that variations observed in XRAIS resolution were indeed due to changes in diffraction quality or anisotropy (Figure 9). These synchrotron data represent a large array of experimental conditions, and not every condition was captured by our forward model. The challenge in creating a generalized resolution prediction model is in preparing the training data, and ensuring it covers the most important scenarios, something still under investigation.

It's worth revisiting the relationship between the B-factor and resolution that XRAIS is based on. Essentially, the XRAIS model is inferring B-factors, and converting them to resolutions via the relationship shown in Figure 1. This relationship is an approximate one (Holton and Frankel 2010), and it is useful to show a comparison between the XRAIS B-factors and those derived from refinement (Figure 10). The XRAIS B-factor of a dataset was computed as

follows: for each diffraction image, XRAIS was used to infer 4 B-factors (one per quadrant). B_{\min} was defined as the minimum B-factor amongst an image's quadrants, and then B_{\min} was averaged across the dataset to obtain the "XRAIS B" shown in Figure 16. We found this correlated best with the Wilson B-factor (Wilson 1942), and the median atomic B-factor refined using *Phenix* (Liebschner et al 2019).

3.3. Multi-lattice detection in diffraction collected at LCLS

Fixed-target diffraction data for reduced [FeFe]-hydrogenase I from *Clostridium pasteurianum* (CpI) crystals were collected at 100 Kelvin using a goniometer-based setup (Cohen et al 2014) at the X-ray pump probe (XPP) hutch of LCLS (Chollet et al 2015), and the results have already been published (Artz et al 2020). Diffraction images from 1 - 3mm crystals were acquired using a helical data collection mode, with unattenuated SASE pulses approximately 40 fs in duration, centered at 9460 eV, and focused into a roughly 20 μm wide interaction region. Crystals were translated by 70 μm and rotated between exposures, resulting in 512 diffraction images. By eye, many of the images were indicative of crystal splitting and this complicated every aspect of the analysis, as originally reported in (Artz et al 2020).

For the results reported in (Artz et al 2020), the images were visually inspected and 122 of the 512 images were manually selected for processing according to their being characteristic of single lattice scattering. As shown in Figure 11, all but 4 of these images were also deemed single lattices according to XRAIS multi-lattice detection. Examples of images flagged as having single/multiple lattices are shown in Figure 12, and the results make sense, intuitively. To seek a more quantitative result, the original data were reprocessed, and with recent software updates to the program *dials.stills_process*, 431 out of the original 512 images were indexed and integrated, and indexed images that had fewer than 120 indexed strong (observed) Bragg reflections were rejected. The remaining images were sorted according to their XRAIS multi-lattice predictions (the values shown in Figure 11) and split into two datasets (Set A and Set B) of 198 images each. Set A consisted of images with average multi-lattice probability of 0.598% (determined from the raw XRAIS multi-lattice predictions), and Set B consisted of images with multi-lattice probability of 41.6% on average. Both sets were merged, and the resulting CC1/2 statistics are shown in Figure 13. Notably, the CC1/2 was lower at wider scattering angles for the set that included more multiple lattice diffraction (Set B). This is in-line with the general assumption that multi-lattice scattering is problematic for most data processing software.

3.4. Multi-lattice detection in diffraction from SSRL

XRAIS multi-lattice detection was performed on the datasets outlined in Table 3. The results are summarized in Figure 14. From these results it was concluded that 4 of the datasets (B, H, N, Y) had a majority amount of multi-lattice scattering ($> 50\%$). A closer look at images from those datasets revealed features indicative of multi-lattice scattering, as shown in Figure 15. Notably, the multi-lattice detection model was trained on simulated images in the Rayonix 345 detector format used during the XPP data collection discussed above, however it worked well on these SSRL datasets consisting PILATUS 6M and EIGER 16M images. This seems to indicate that the multi-lattice features the model looks are related to the Bragg peak profiles, and mostly independent of the underlying detector geometry. One complication appears to be ice and salt Bragg peaks being mistaken for protein crystal Bragg peaks. However, the presence of ice and salt diffraction is still useful information to know when collecting data, and these features can always be masked.

3.5. Implementation in Interceptor

Both the resolution prediction and multi-lattice detection models are implemented in the live X-ray diffraction image analysis program *Interceptor*, developed in-house at the Stanford Synchrotron Radiation Lightsource (SSRL) (Figure 16). *Interceptor*, running continuously at four crystallography beamlines at SSRL for the past two years, is designed to 1) balance the load among many distributed image-analysis workers, 2) minimize disk I/O, 3) handle scenarios of worker shortage at peak capacity, and 4) provide the workers immediate access to individual images even before they are incorporated into an aggregate file format like HDF5. The architecture is implemented using the ZeroMQ messaging library, with available workers requesting images using the ZeroMQ REQ protocol and the data collection software replying with the REP protocol after each image is written to disk. The original version of Interceptor relied on diffraction image processing algorithms implemented in *DIALS* (Winter et al 2018). In the current version, XRAIS models are utilized concurrently with the conventional image processing algorithms.

3.6. Processing times on GPU

We carefully timed XRAIS resolution modeling using a 24 core (Intel Xeon Gold 6126 2.6 GHz) machine running Centos 7 with an Nvidia A100 GPU. The GPU was utilized by multiple cores in parallel, and parallelization was done by evenly dividing diffraction images over cores using the Message Passing Interface (MPI) protocol. The results are shown in

Table 4. Depending on the number of quadrants used, the detector model, and the number of cores running simultaneously, the per-image downsampling times ranged from 7 - 70 ms, and the per-image inference times ranged from 7.4 - 22 ms. With this one machine, using all 24 ranks, the A100, and only using 1 quadrant for inference, EIGER 16M images were processed at 97.7 Hz and PILATUS 6M images at 261 Hz, including time taken to read the images from disk using the FabIO library. Of course, these times will vary depending on the way raw pixels are handled in disk and RAM, and whether detectors must first write to disk before moving data to processing machines. Without the GPU, these processing rates fell to 18.1 Hz (EIGER 16M) and 20.4 Hz (PILATUS 6M). While *Interceptor* is routinely run on CPU-only machines, it utilizes a larger network of multiple machines to boost throughput. These results, however, suggest that GPU machines have great potential for providing faster real-time feedback to users.

3.7. Quadrant variation

Due to the timing test results shown in Table 4, for high frame rate experiments, it may sometimes be beneficial to use a single quadrant for inference. Indeed, this was done for all the multi-lattice results shown in this report, however for the resolution inference results shown in Figures 8-10, all 4 image quadrants were used to infer resolution separately, and then the mean (or minimum) was taken as the effective resolution. Looking at the entire image to gauge its resolution is perhaps the most accurate approach, but it is instructive to explore the variation in resolution with each quadrant. This is shown in Figure 17 for the 25 SSRL datasets from Table 3 and Figure 8. In most of the datasets, the resolutions are similar regardless of quadrant, however anisotropic diffraction and issues with beam centering can influence per-quad resolution (e.g., Figure 17A). Future versions of XRAIS will train on more diverse datasets to yield even more precise resolution estimates and will perhaps recognize (from a single quadrant) when the scattering is asymmetric or when the incident beam is misaligned.

4. Discussion

AI as a tool is inherently tied to automation. The central utility of computers is to enhance the human experience by automating routine tasks, and this goes for crystallographers as well. Indeed, data analysis at SR crystallography beamlines have increasingly become automated (Cornaciu et al 2021, Douangamath et al 2021, Tsai et al 2013) and this is also true for XFEL experiments. For example, during two recent LCLS experiments targeting small molecules

and viral COVID 19 proteins (Blaschke et al 2021), data were recorded at SLAC and automatically transferred using XROOT protocol to the National Energy Research Scientific Computing Center (NERSC) for high-performance computing. Data processing jobs were submitted to NERSC compute nodes remotely by the *cctbx.xfel* application (Brewster et al 2019), and preliminary structure solutions were automatically uploaded to a web server for experimenter assessment, in as little as ten minutes after 120 Hz data collection began. This did require initial user inputs for indexing, integration, merging, and structure refinement, but with the addition of new AI programs (Ke et al 2018, Rahmani et al 2023) to screen for diffraction, and our present body of work that uses AI to characterize diffraction, we are moving closer to total machine control of our analysis.

One drawback of using these supervised learning approaches is the sensitivity to training data content. Indeed in (Ke et al 2018 and Rahmani et al 2023), the authors concluded that their training datasets did not readily adapt to new data from different detectors or to data collected under different conditions. We have seen training set bias in our own work as well. The benefit of our simulation-to-model approach is that the simulations are fully within our control, allowing us to readily expand training datasets to adjust for shortcomings and to adapt to new experimental setups. Further, with the XRAIS framework, we are well positioned to begin exploring the prediction of other interesting experimental parameters. We are actively exploring using XRAIS models to determine the “beam center” on a diffraction pattern, and the preliminary results are encouraging. These models could then be used to warn users when the geometry is becoming misaligned. In addition to providing real-time feedback, we expect XRAIS will reduce the time and effort required to aptly process challenging datasets. XRAIS can detect and flag problematic diffraction images, such as images with split diffraction spots, which are difficult to identified with the currently available processing software. Inclusion of problematic diffraction images reduces the quality of merged data, making structure determination (especially by *ab initio* phasing) difficult or impossible. We expect XRAIS will be the key to identifying which diffraction images should be included and processed to yield usable merged datasets.

Already AI can scale and merge structure factor intensity measurements in serial and rotation crystallography (Dalton et al 2022). AI tools like XRAIS can potentially be used to inform users of progress towards full datasets, or how to adjust beamline parameters to optimize chances for experimental success. We are moving towards a new era when we can ask the AI to estimate unit cell constants from single images, and index and integrate those images

automatically. The possibilities are exciting, and we hope this paper serves as inspiration for the community to determine how best to apply these tools to better aid the crystallographic community, and ultimately to lead to better results and higher throughput experiments at crystallography beamlines in general.

Figure 1 Resolution versus B-factor. Structures from the PDB indicate an overall trend in resolution vs B-factor, as determined by a simple quadratic fit (red line). The square markers shown here represent average B-factors at each resolution across the entire PDB, and the error bar is one standard deviation. The fit was only done using data on the interval (1 Å - 4.5 Å). This trend becomes unreliable at resolutions > 5.5 Å.

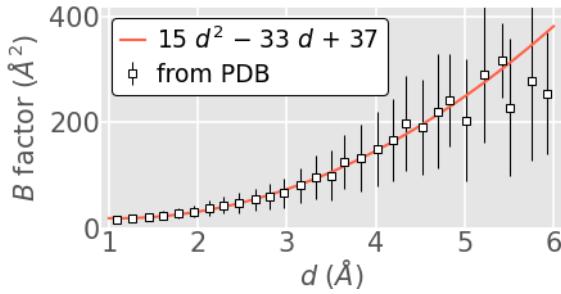


Figure 2 Simulated PILATUS 6M images with varying resolutions. Some parameters underlying each image are summarized in Table 1. For each sub-image, the beam center is in the upper-left corner and the resolution (determined from the B-factor) is indicated by a red-dashed line. The sub-images represent one quadrant of a PILATUS 6M camera, down sampled to a 512x512 pixel array according to Section 2.2 (see also Figure 4). Note, resolution here is related to the B-factor by the relationship shown in Figure 1. Hence, while sometimes the resolution appears intuitively as the point where the scattering drops off (e.g., in A, B, C, D, G, H, I, K, and O), other times the Bragg reflections extend to wider angles, beyond the indicated resolution (E, J, L, and N). This results from inaccuracy in the resolution-versus-B-factor relationship (Figure 1). Rarely, resolution is obscured by large background (F) or is covered by the beam stop (M). These edge cases add noise to the model training. Colorbars are shown in the bottom left in square-root-photon units.

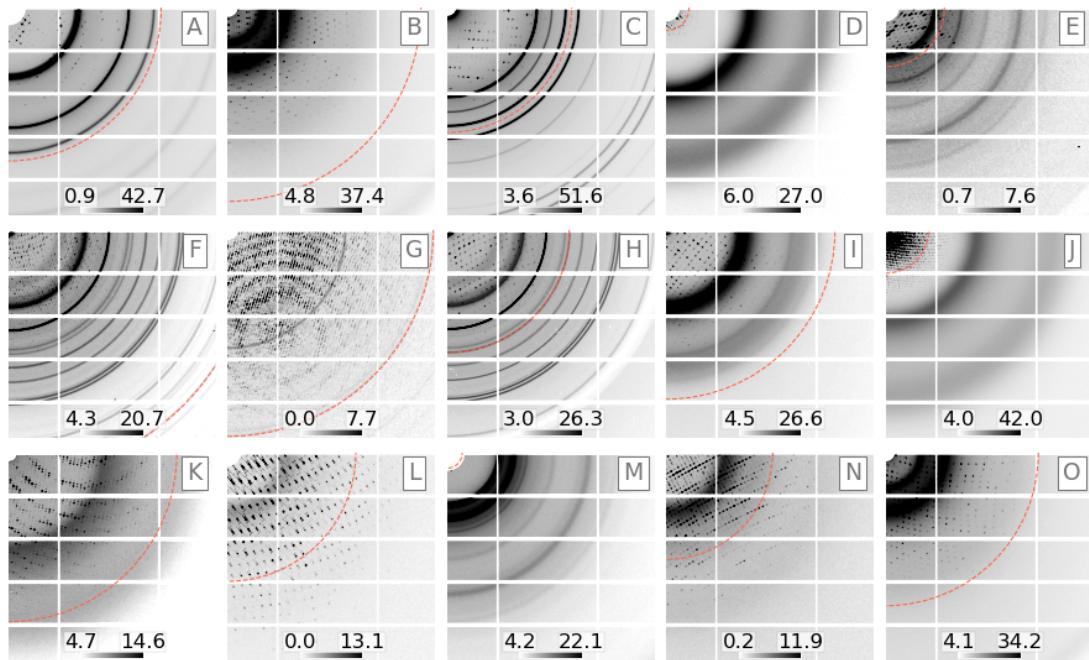


Figure 3 Simulated diffraction from 1 lattice (top row), 2 lattices (middle row), or 3 lattices (bottom row). Each subimage represents the lower quadrant of a Rayonix camera, downsampled as illustrated by Figure 4. Colorbars are shown in the lower-right corner in square-root-photon units.

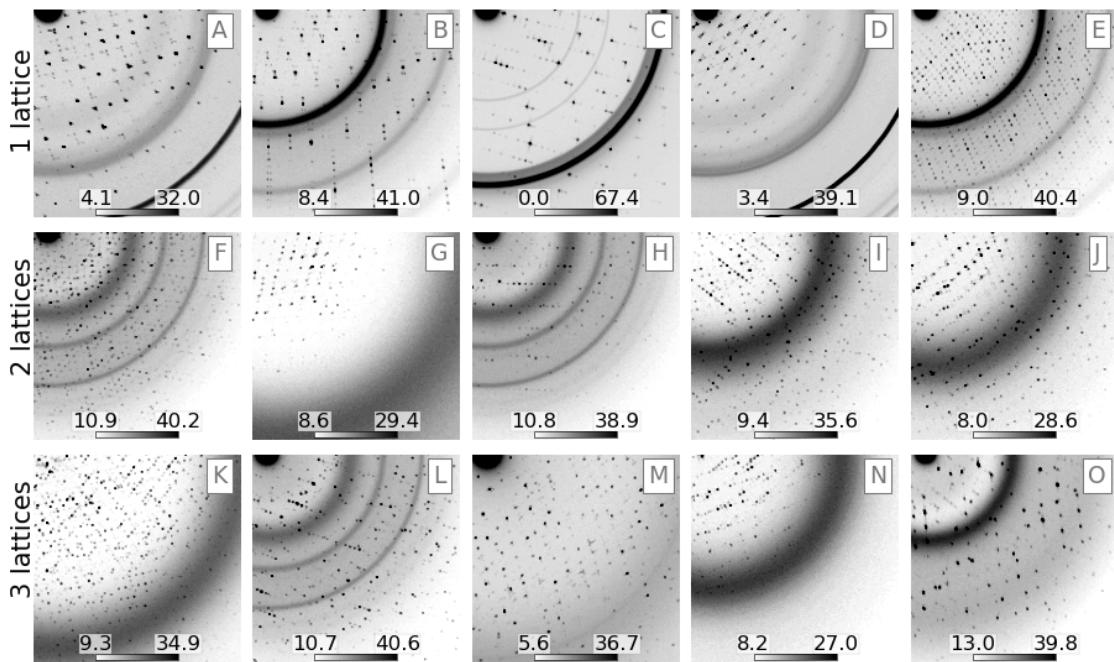


Figure 4 Downsampling scheme. Shown is a region of a PILATUS 6M image with a Bragg reflection (with numbers indicating pixel value). Raw data (left) are divided into blocks of pixels (indicated by red dashed lines). This occurs for both simulated and experimental data. The “conditioned” pixel value (right) is the square root of the maximum pixel value within each block, cast to an integer (floor operation). Block size varies according to detector model, either 2x2 blocks (PILATUS 6M) or 4x4 blocks (EIGER 16M, JUNGFRAU16M, Rayonix 345) were used.

26	35	26	28	34	30	24	24		
33	29	19	29	38	35	37	28		
27	18	34	30	34	37	19	37		
35	34	127	194	74	30	30	27		
26	29	63	318	298	66	36	27		
27	28	30	70	101	38	37	29		
25	36	49	36	34	50	26	30		
26	22	34	26	23	37	31	27		

5	5	6	6	
5	13	8	6	
5	17	17	6	
6	7	7	5	

Figure 5 Model architecture. (A) Raw data were downsampled as described in Section 2.2, forming four 512x512 quadrants. Quadrants were then passed through a ResNet architecture, resulting in 1000 features. Next, a series of fully connected layers (FC1, FC2) was used to convert the 1000 features into a scalar value. If predicting resolution (B), this was converted to an inverse resolution using the diffraction wavelength (λ), down-sampled pixel size (p), and sample-to-detector distance (D). If predicting multi-lattice scattering (C), this scalar was passed through a Sigmoid function and then rounded, such that 0, 1 indicated single, multiple lattice scattering, respectively. The image and line plots in (A) are from a real experimental image as it was passed through the fully trained resolution model. The inferred resolution in this case was 1.67 Å. Table 2 describes the number of parameters in the different model stages (ResNet, FC1, FC2). One quadrant was enough to predict quantities of interest, however repeated model passes with the 2nd, 3rd and 4th quadrants can provide a measure of uncertainty in the predicted values.

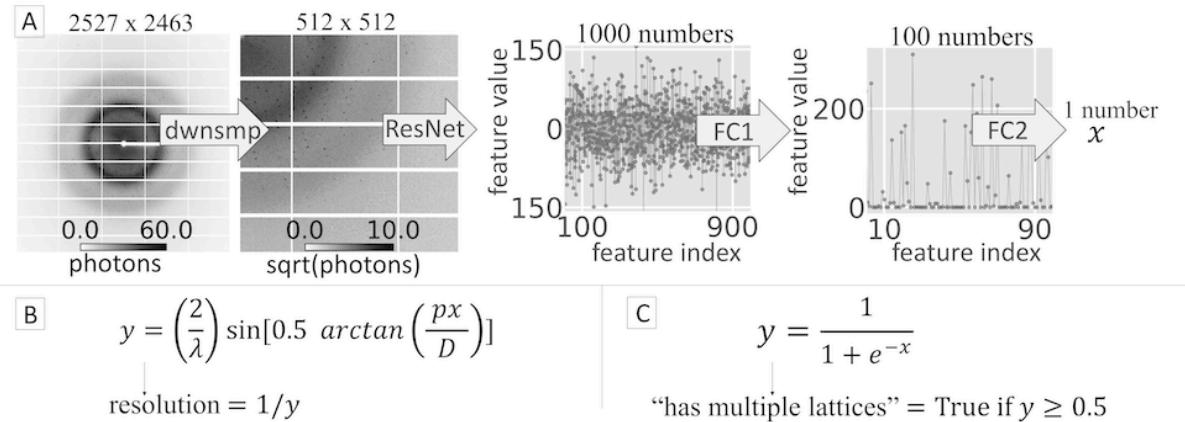


Figure 6 Training optimization curves. (A) Accuracy versus training epoch for the resolution prediction model. This is a regression model, for which we define accuracy as the fraction of images

whose predictions are within 0.07 inverse Angstroms of the ground truth. The training job was stopped after epoch 354 and then restarted, as indicated by the discontinuity. (B) Accuracy versus epoch for the multi-lattice detection model. Here, accuracy is the fraction of predictions with the correct label. For both plots, the test curves (black markers) are derived from images never used for training. Eventually, training accuracy diverges, indicating model bias. The vertical lines mark the epoch where we saved the models for use with experimental data.

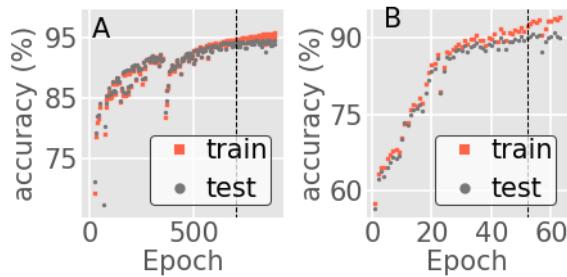


Figure 7 Resolution prediction for JUNGFRAU 16M data collected at the SwissFEL light source. All colorbars are in square root photon units. (A) A quadrant of the JUNGFRAU16M image (512 x 512 pixels) with the highest predicted resolution (1.3 Å) (B) Same as (A) but zoomed-in to the outer corner of the image, showing high resolution Bragg peaks. (C) CC1/2 versus resolution for 3 different merges. Images were sorted according to XRAIS predicted resolution and then divided into 3 groups called “high”, “mid”, and “low”. The high, mid, low groups included images whose XRAIS predicted resolutions lay in the range (1.3 - 2.5 Å), (2.0 - 2.9 Å), (2.5 - 5.2 Å), respectively. The images in each group were processed with *dials.stills_process* and merged with *cctbx.xfel.merge*, resulting in 3 CC1/2 curves. As shown, images with higher resolutions yielded better CC1/2 statistics and indicated that the model can be used to accurately sort images based on resolution (D) Histogram of the resolutions predicted for all 9,592 images. (I-IV) Maximum composite images. A maximum composite image is an image whose pixel value is the maximum across a subset of images (see Brewster et al 2015 for further description). In this case, the subsets are those images whose resolutions fell within the shaded regions in (D). I, II, III, and IV correspond to intervals (1.3 - 1.6 Å), (2.0 - 2.2 Å), (2.7 - 2.8 Å), and (3.8 - 5.7 Å) and have 640, 640, 639, 639 images each, respectively. The average resolution in each max composite image is labeled by a black, dashed circle. The large peaks in these images are from salt crystals or other parasitic scatterers in the beam, and the smaller, more densely packed peaks represent Bragg reflections.

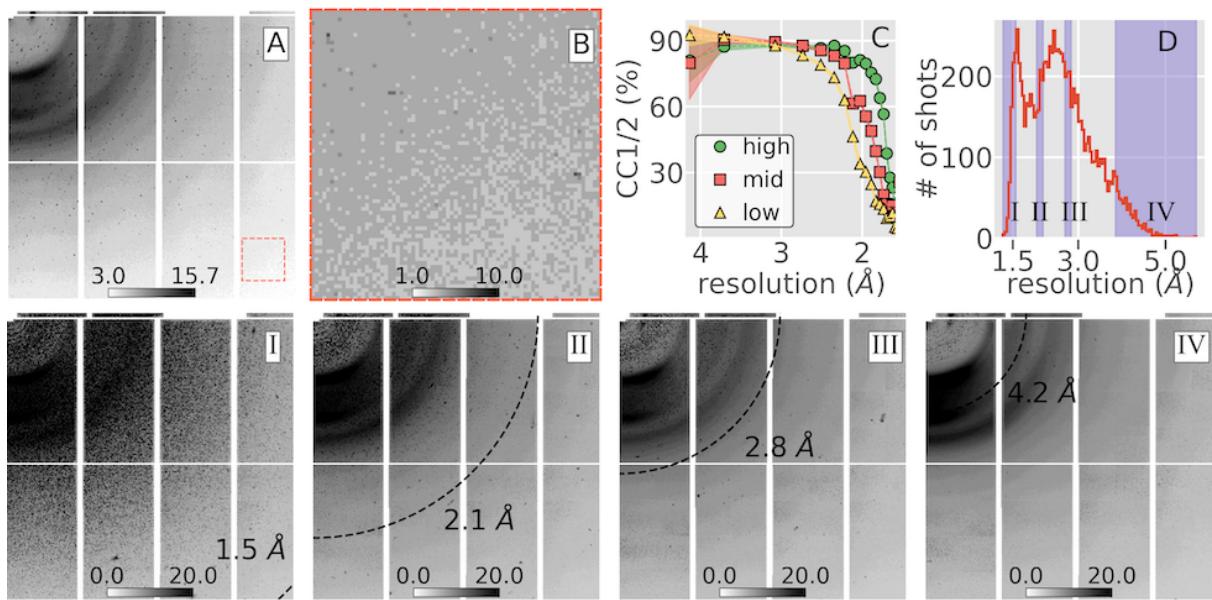


Figure 8 XRAIS resolution prediction for rotational datasets recorded at the SSRL Structural and Molecular Biology beamlines. The horizontal axis is sorted according to collection time. The red, blue markers are the minimum, average inferred resolution per shot (across the 4 quadrants), and the gray markers are resolution estimates from *DIALS*. Each dataset is labeled by a resolution determined using Aimless (shown at the top of each subplot and indicated by the dashed line) as the point where the overall signal-to-noise ratio of the integrated intensity dipped below 1.5. The shaded regions labeled I, II in subplots A, S, and W correspond to regions where maximum composite images were computed to identify the cause of systematic variation in the XRAIS-predicted resolution (see Figure 9 for the corresponding maximum composite images).

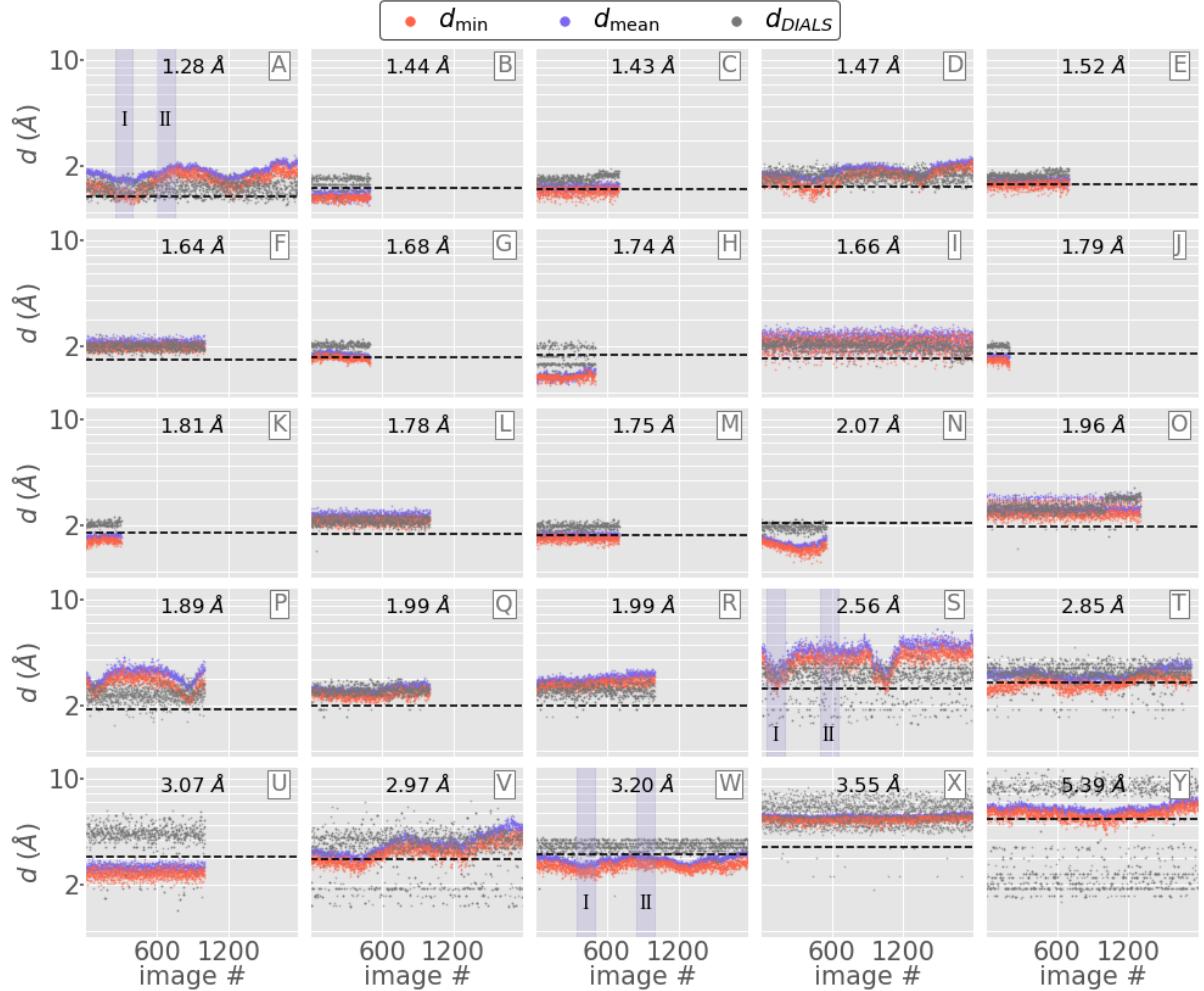


Figure 9 Max-composite images of sequential groups of SSRL diffraction patterns. Resolution labels 3.20 Å, 2.56 Å, and 1.28 Å correspond to the SSRL datasets that are shown in Figure 8 (subplots W, S, and A, respectively). The shot ranges over which the maximum composite images were computed are shown in the subplots of Figure 8 (W, S, and A). Regions labeled by “I” correspond to relatively high resolution estimates, and regions labeled by “II” correspond to low resolution estimates from the same dataset. The black circles indicate the nominal resolutions of the datasets (taken from the Aimless logs), and the red circles mark the average resolutions determined by XRAIS for the images that went into each maximum composite. Different features in the data can influence XRAIS resolution. Here we observe asymmetric diffraction influencing the predicted resolution of the pattern.

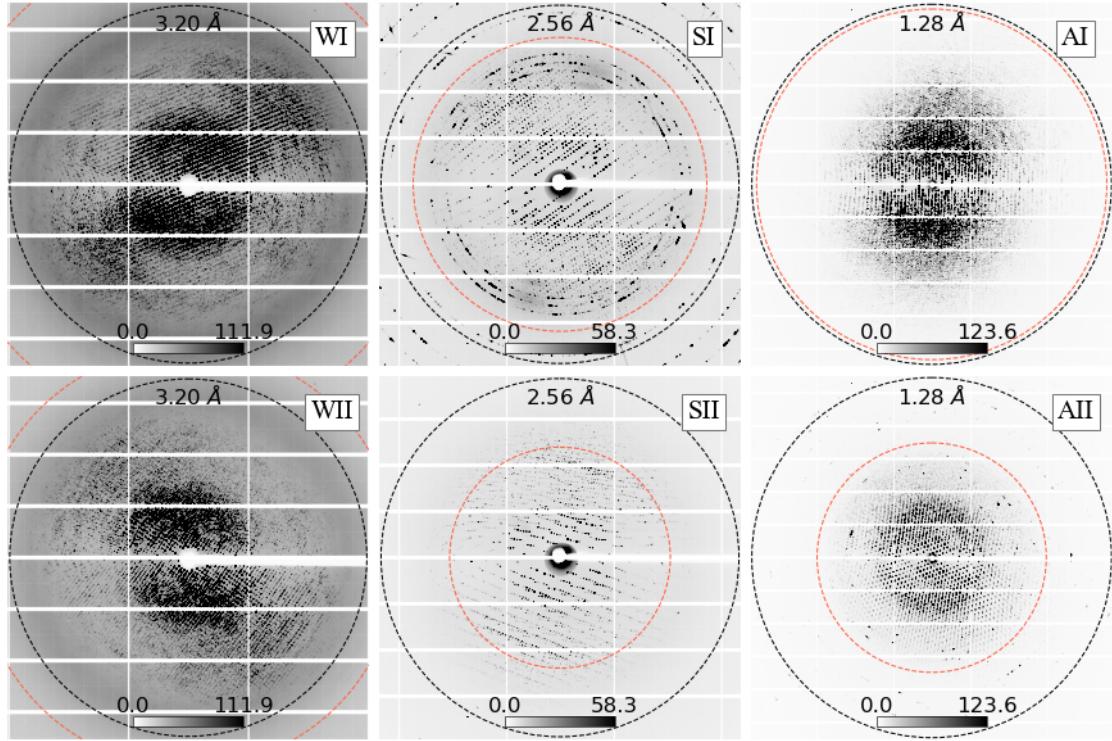


Figure 10 B-factor inference. Overall B-factors and resolutions were inferred using XRAIS and compared to the Wilson B-factor and the refined atomic B-factors, for each of the SSRL datasets shown in Table 3 and Figure 8. (Top) The XRAIS B-factor, the Wilson B factor (from *Aimless*), and the median atomic B-factor (from structure solution and refinement using *Phenix*). (Bottom) The overall resolution of each dataset, defined as the average over the per-image quantity d_{\min} , where d_{\min} is the lowest of 4 resolutions, one per quadrant.

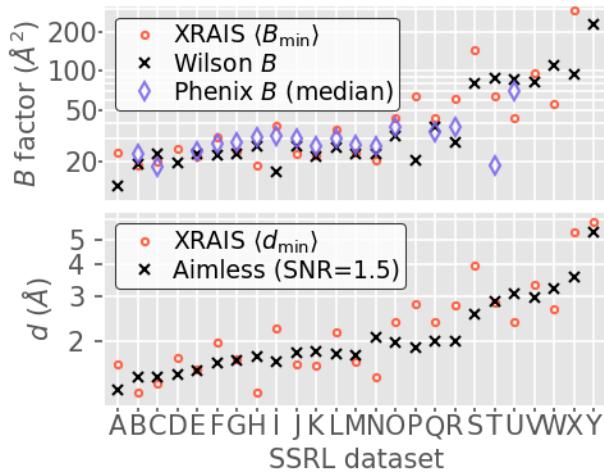


Figure 11 XRAIS multi-lattice detection for fixed-target data collected at LCLS (Artz et al 2020). The raw prediction corresponds to the probability that the image contained multi-lattice diffraction, hence subtracting this number from 1 computes the probability that the image only contained diffraction from a single lattice. The gray histogram represents all 512 images, whereas the

red histogram represents the 122 images that were hand-selected for processing in (Artz et al 2020). These hand-selected images were chosen because they resembled good quality, single-lattice diffraction. Only 4 of these hand-selected images were predicted by XRAIS to contain multi-lattice scattering (with multi-lattice probability > 50%).

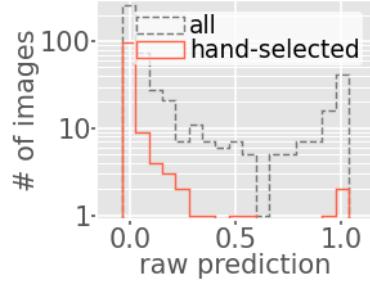


Figure 12 Comparing multi- and single- lattice diffraction from the XPP dataset. Images A-C were flagged by XRAIS as containing multi-lattice diffraction, and images D-F were flagged as containing diffraction from a single lattice. It is obvious visually that images A-C contain more disordered diffraction, indicative of multi-lattice scattering. Colorbars are in square root photon units.

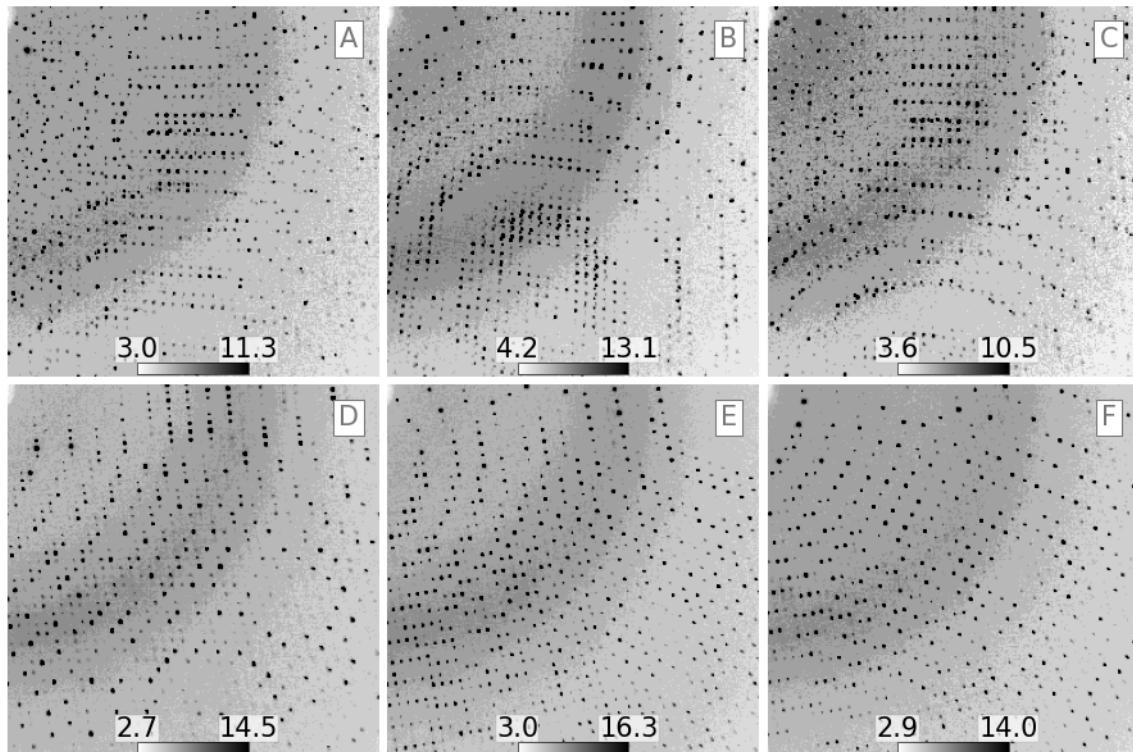


Figure 13 Merge statistics for hydrogenase data collected at XPP. After indexing the images, they were grouped into two sets according to each image's probability of containing multi-lattice diffraction (according to XRAIS). Each set contained 198 images. (A) The CC1/2 obtained after merging either set using *cctbx.xfel.merge*. CC1/2 was computed 5 times per set with random half-dataset assignments. The markers represent the mean, and the shaded region indicates ± 1 standard

deviation from that mean. (B) The merged multiplicity in the asymmetric unit. Notably, Set B had a slightly higher overall multiplicity (7.83 vs 7.36), but a lower CC1/2 at high resolution.

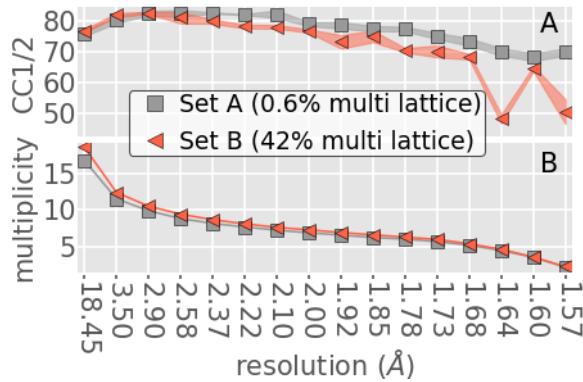


Figure 14 XRAIS multi-lattice detection in the 25 SSRL datasets described in Table 3 and shown in Figure 8. The y-axis here indicates the probability that an image from the dataset contained multi-lattice diffraction.

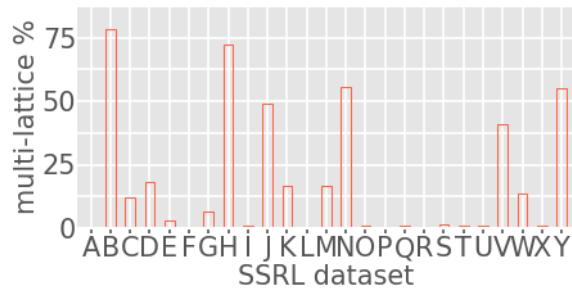


Figure 15 Maximum composite images for the first 20 exposures (4° total rotation) from the SSRL datasets indicated by the subplot labels B, H, N, D, I, and O. Based on XRAIS multi-lattice detection (see Figure 14) it was known that datasets B, H, and N had a high chance of containing multi-lattice scattering (80%, 74%, 57%, respectively). On the contrary, datasets D, I and O had a low chance of containing multi-lattice scattering (12%, 0%, 0%, respectively). All colorbars are in “square root photon” units. While dataset N clearly shows multi-lattice features from separate protein crystals, it appears datasets B and H contain more subtle features (streaks and ice Bragg peaks) that caused XRAIS to infer the presence of multiple lattices.

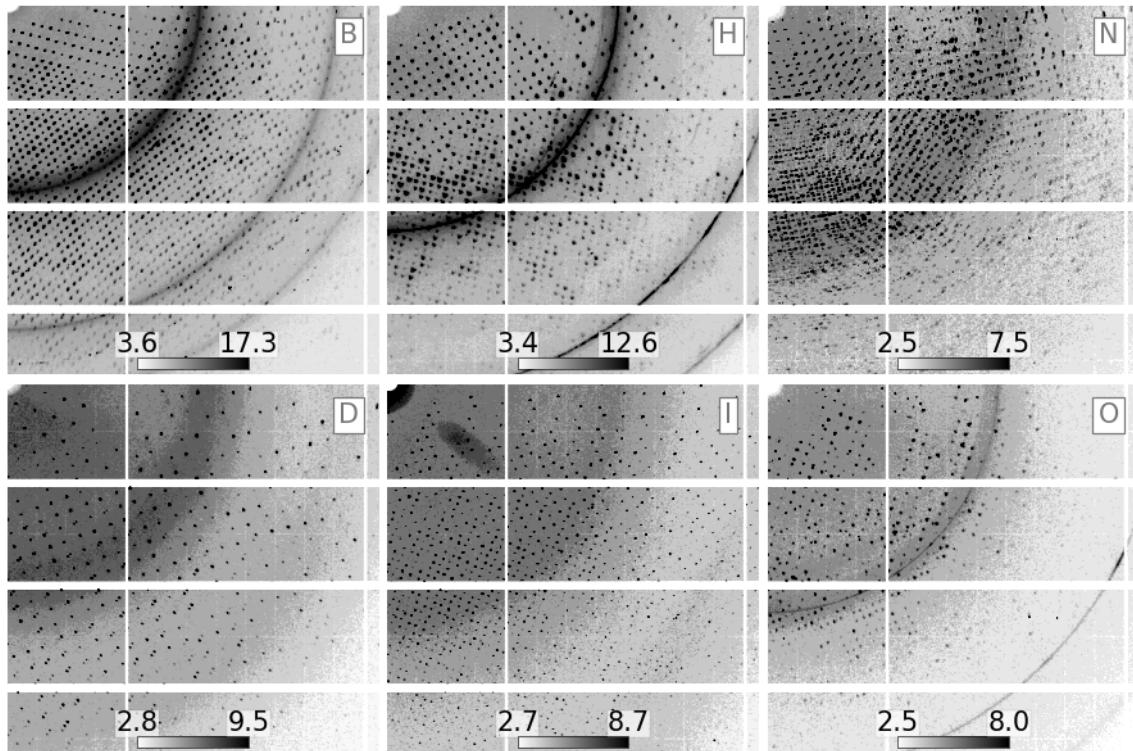


Figure 16 Live X-ray diffraction image analysis with *Interceptor*. (A) The images are recorded on the filesystem and subsequently forwarded to a set of parallel processing modules. The results are then forwarded to the *BluIce* beamline control software. (B) Users can visualize the results in a configurable strip chart, which is updated as the data are collected. The configuration shown consists of a plot of resolution (top, red) and multi-lattice probability (bottom, black/grey) vs. image number.

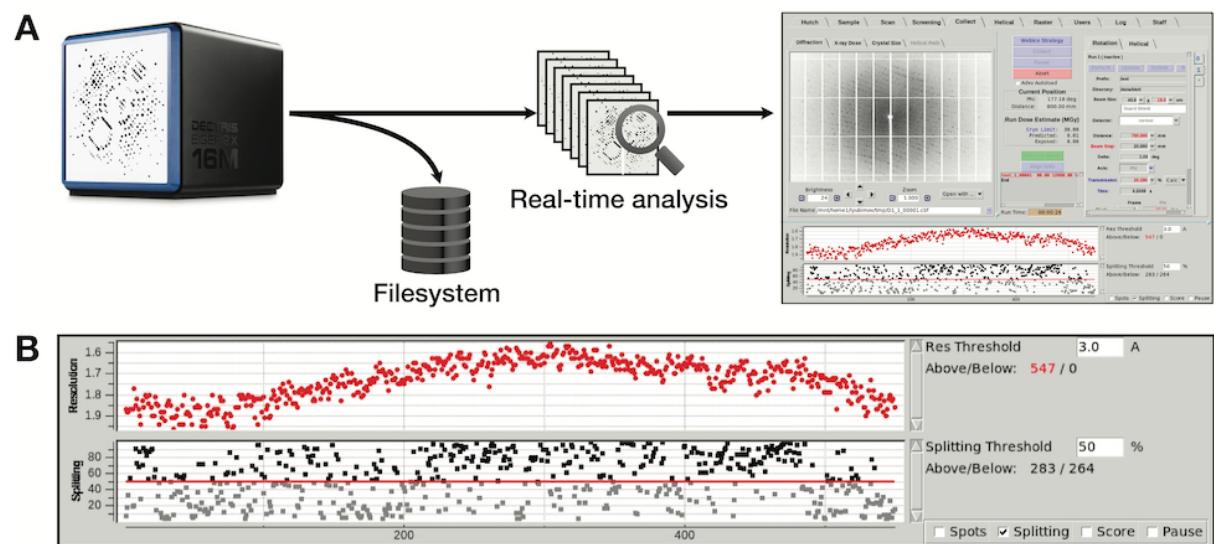


Figure 17 Same as Figure 8, however separate inferences are shown for each quad.

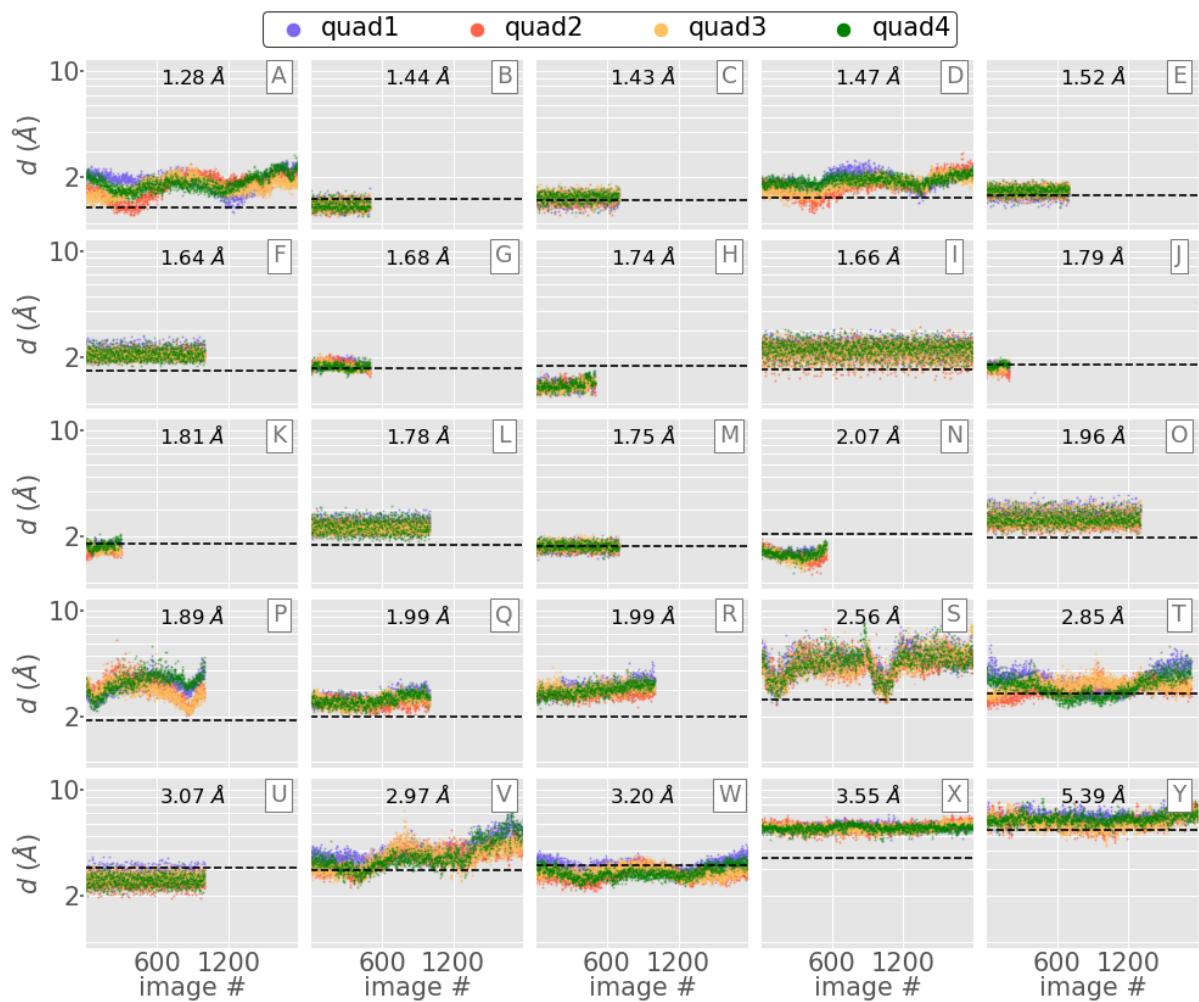


Table 1 Simulation properties corresponding to Figure 2.

Figure 1 Label	Resolution (Å)	Distance (mm)	Background scale
A	2.06	217	1.25
B	1.71	221	1.25
C	2.46	240	1.25
D	12.8	253	1
E	5	224	0.05
F	1.43	253	1
G	1.63	298	0.01
H	2.5	250	1
I	1.9	225	1

J	6.82	274	1.25
K	1.9	293	1.25
L	2.4	260	0.02
M	19.2	238	1
N	2.84	264	0.1
O	2.07	222	1

Table 2 Model fitting details.

Predictor type	resolution	splitting
Resnet parameters	25,550,760	21,791,400
Total parameters	25,650,961	21,891,601
optimizer	SGD	SGD
FC1 dropout	No	Yes
Training images	201,000	117,000
Momentum	0.9	0.9983
Weight decay	N/A	2.5e-4
Loss function	MAE	BCE
Learning rate	6e-3	1.04e-3

SGD: sparse gradient descent, MAE: mean absolute error, BCE: binary cross entropy.

Table 3 SSRL crystallography datasets tested with XRAIS.

Dataset ID	Aimless resolution (Å)	Distance (mm)	Detector
A	1.28	200	PILATUS 6M
B	1.44	250	PILATUS 6M
C	1.43	250	PILATUS 6M
D	1.47	250	PILATUS 6M

E	1.52	250	PILATUS 6M
F	1.64	300	PILATUS 6M
G	1.68	300	PILATUS 6M
H	1.74	300	PILATUS 6M
I	1.66	300	PILATUS 6M
J	1.79	300	PILATUS 6M
K	1.81	300	PILATUS 6M
L	1.78	300	PILATUS 6M
M	1.75	250	PILATUS 6M
N	2.07	300	PILATUS 6M
O	1.96	300	PILATUS 6M
P	1.89	275	EIGER 16M
Q	1.99	300	PILATUS 6M
R	1.99	300	PILATUS 6M
S	2.56	350	PILATUS 6M
T	2.85	300	PILATUS 6M
U	3.07	300	PILATUS 6M
V	2.97	350	PILATUS 6M
W	3.20	400	PILATUS 6M
X	3.55	350	EIGER 16M
Y	5.39	400	PILATUS 6M

Table 4 GPU time tests on a single Nvidia A100 GPU.

Processor cores were used in parallel, each sharing the A100.

Parallel cores	Number of quadrants	Detector	Downsample (ms)	Inference (ms)	Downsample CPU (ms)	Inference CPU (ms)
8	1	EIGER 16M	22 (0.9)	8.0 (1.3)	-	-
8	4	EIGER 16M	70 (4.5)	9.2 (3.3)	-	-

24	1	EIGER 16M	42 (5.8)	12 (3.9)	61 (6.1)	1010 (63)
8	1	PILATUS 6M	7.1 (0.4)	7.4 (0.6)	-	-
8	4	PILATUS 6M	18 (1.5)	9.6 (4.3)	-	-
24	1	PILATUS 6M	13 (2.8)	22 (8.5)	22 (1.9)	1084 (54)

Downsample times are medians over SSRL Datasets P, D for EIGER 16M, PILATUS 6M, respectively. CPU-only times are shown for reference.

Acknowledgements We acknowledge the Paul Scherrer Institute, Villigen, Switzerland for provision of free-electron laser beamtime at the Bernina instrument of the SwissFEL ARAMIS/ATHOS branch and we thank Nick Sauter, Jan Kern, Aimin Liu, and Romie Nguyen for use of the Cytochrome measurements collected there. Use of the Stanford Synchrotron Radiation Lightsource, SLAC National Accelerator Laboratory, is supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. DE-AC02-76SF00515. The SSRL Structural Molecular Biology Program is supported by the DOE Office of Biological and Environmental Research, and by the National Institutes of Health, National Institute of General Medical Sciences (P30GM133894). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of NIGMS or NIH. This research used resources of the National Energy Research Scientific Computing Center (NERSC), a U.S. Department of Energy Office of Science User Facility located at Lawrence Berkeley National Laboratory, operated under Contract No. DE-AC02-05CH11231 using NERSC award BER-ERCAP0024756. D Mendez and his contributions to this project were supported by the Department of Energy, Laboratory Directed Research and Development program at SLAC National Accelerator Laboratory, under contract **DE-AC02-76SF00515**. D Mendez thanks Chris Young for productive discussions, and Yuer Hao and the UC Santa Barbara Capstone Data Science Initiative for a fruitful collaboration.

References

- Allahgholi, A., Becker, J., Delfs, A., Dinapoli, R., Göttlicher, P., Graafsma, H., Greiffenberg, D., Hirsemann, H., Jack, S., Klyuev, A., Krüger, H., Kuhn, M., Laurus, T., Marras, A., Mezza, D., Mozzanica, A., Poehlsen, J., Shefer Shalev, O., Sheviakov, I., ... Zimmer, M. (2019). Megapixels @ Megahertz – The AGIPD high-speed cameras for the European XFEL. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 942, 162324. <https://doi.org/10.1016/J.NIMA.2019.06.065>

- Antipov, S. P., Assoufid, L., Grizolli, W. C., Qian, J., & Shi, X. (2018). The LCLS-II: A High Power Upgrade to the LCLS. *Proc. 9th International Particle Accelerator Conference (IPAC'18)*, 18–23. <https://doi.org/10.18429/JACOW-IPAC2018-MOYGB2>

Artz, J. H., Zadvornyy, O. A., Mulder, D. W., Keable, S. M., Cohen, A. E., Ratzloff, M. W., Williams, S. G., Ginovska, B., Kumar, N., Song, J., McPhillips, S. E., Davidson, C. M., Lyubimov, A. Y., Pence, N., Schut, G. J., Jones, A. K., Soltis, S. M., Adams, M. W. W., Raugei, S., ... Peters, J. W. (2020). Tuning Catalytic Bias of Hydrogen Gas Producing Hydrogenases. *Journal of the American Chemical Society*, 142(3), 1227–1235.

https://doi.org/10.1021/JACS.9B08756/ASSET/IMAGES/LARGE/JA9B08756_0005.jpeg

Baxter, E. L., Aguila, L., Alonso-Mori, R., Barnes, C. O., Bonagura, C. A., Brehmer, W., Brunger, A. T., Calero, G., Caradoc-Davies, T. T., Chatterjee, R., Degrado, W. F., Fraser, J. S., Ibrahim, M., Kern, J., Kobilka, B. K., Kruse, A. C., Larsson, K. M., Lemke, H. T., Lyubimov, A. Y., ... Cohen, A. E. (2016). High-density grids for efficient data collection from multiple crystals. *Acta Crystallographica Section D: Structural Biology*, 72(1), 2–11.

<https://doi.org/10.1107/S2059798315020847/YT5086SUP3.mov>

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28(1), 235–242.

<https://doi.org/10.1093/NAR/28.1.235>

Blaschke, J. P., Brewster, A. S., Paley, D. W., Mendez, D., Sauter, N. K., Kröger, W., Shankar, M., Enders, B., & Bard, D. (2021). Real-Time XFEL Data Analysis at SLAC and NERSC: a Trial Run of Nascent Exascale Experimental Data Analysis. *ArXiv*. <https://arxiv.org/abs/2106.11469v2>

Bragg, W. H. (1914). XCVII. The intensity of reflexion of X rays by crystals . *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 27(161), 881–899.
<https://doi.org/10.1080/14786440508635159>

Brändén, G., & Neutze, R. (2021). Advances and challenges in time-resolved macromolecular crystallography. *Science*, 373(6558).

<https://doi.org/10.1126/SCIENCE.ABA0954/ASSET/4C9C3E09-F268-4D7B-9E4E-F70396F031FD/ASSETS/IMAGES/LARGE/SCIENCE.ABA0954-F6.JPG>

Brewster, A. S., Young, I. D., Lyubimov, A., Bhowmick, A., & Sauter, N. K. (2019). Processing serial crystallographic data from XFELs or synchrotrons using the cctbx.xfel GUI. *Computational Crystallography Newsletter*, 10, 22–39.

<http://portal.nersc.gov/archive/home/projects/cxidb/www/43/HAD13a/run371999-0.h5>

Casanas, A., Warshamanager, R., Finke, A. D., Panepucci, E., Olieric, V., Nöll, A., Tampé, R., Brandstetter, S., Förster, A., Mueller, M., Schulze-Briese, C., Bunk, O., & Wang, M. (2016). EIGER detector: Application in macromolecular crystallography: *Acta Crystallographica Section D: Structural Biology*, 72(9), 1036–1048. <https://doi.org/10.1107/S2059798316012304>

Chollet, M., Alonso-Mori, R., Cammarata, M., Damiani, D., Defever, J., Delor, J. T., Feng, Y., Glownia, J. M., Langton, J. B., Nelson, S., Ramsey, K., Robert, A., Sikorski, M., Song, S., Stefanescu, D., Srinivasan, V., Zhu, D., Lemke, H. T., & Fritz, D. M. (2015). The X-ray Pump-Probe instrument at the Linac Coherent Light Source. *Journal of Synchrotron Radiation*, 22(3), 503–507. <https://doi.org/10.1107/S1600577515005135>

Cohen, A. E., Soltis, S. M., González, A., Aguilera, L., Alonso-Mori, R., Barnes, C. O., Baxter, E. L., Brehmer, W., Brewster, A. S., Brunger, A. T., Calero, G., Chang, J. F., Chollet, M., Ehrensberger, P., Eriksson, T. L., Feng, Y., Hattne, J., Hedman, B., Hollenbeck, M., ... Hodgson, K. O. (2014). Goniometer-based femtosecond crystallography with X-ray free electron lasers. *Proceedings of the National Academy of Sciences of the United States of America*, 111(48), 17122–17127. <https://doi.org/10.1073/PNAS.1418733111>

Cohen, A. E. (2021). A new era of synchrotron-enabled macromolecular crystallography. In *Nature Methods* (Vol. 18, Issue 5, pp. 433–434). Nature Research. <https://doi.org/10.1038/s41592-021-01146-y>

Cornaciu, I., Bourgeas, R., Hoffmann, G., Dupeux, F., Humm, A. S., Mariaule, V., Pica, A., Clavel, D., Seroul, G., Murphy, P., & Márquez, J. A. (2021). The Automated Crystallography Pipelines at the EMBL HTX Facility in Grenoble. *JoVE (Journal of Visualized Experiments)*, 2021(172), e62491. <https://doi.org/10.3791/62491>

Dalton, K. M., Greisman, J. B., & Hekstra, D. R. (2022). A unifying Bayesian framework for merging X-ray diffraction data. *Nature Communications*, 13(1), 1–13. <https://doi.org/10.1038/s41467-022-35280-8>

DePonte, D. P., Weierstall, U., Schmidt, K., Warner, J., Starodub, D., Spence, J. C. H., & Doak, R. B. (2008). Gas dynamic virtual nozzle for generation of microscopic droplet streams. *Journal of Physics D: Applied Physics*, 41(19), 195505.

Douangamath, A., Powell, A., Fearon, D., Collins, P. M., Talon, R., Krojer, T., Skyner, R., Branda-Neto, J., Dunnett, L., Dias, A., Aimon, A., Pearce, N. M., Wild, C., Gorrie-Stone, T., & von Delft, F.

(2021). Achieving Efficient Fragment Screening at XChem Facility at Diamond Light Source. *JoVE (Journal of Visualized Experiments)*, 171, e62414. <https://doi.org/10.3791/62414>

Evans, P. R., & Murshudov, G. N. (2013). How good are my data and what is the resolution? *Acta Crystallographica Section D: Biological Crystallography*, 69(7), 1204–1214.

Fielding, A. J., Dornevil, K., Ma, L., Davis, I., & Liu, A. (2017). Probing Ligand Exchange in the P450 Enzyme CYP121 from Mycobacterium tuberculosis: Dynamic Equilibrium of the Distal Heme Ligand as a Function of pH and Temperature. *Journal of the American Chemical Society*, 139(48), 17484–17499. <https://doi.org/10.1021/jacs.7b08911>

Fuller, F. D., Gul, S., Chatterjee, R., Sethe Burgie, E., Young, I. D., Lebrette, H., Srinivas, V., Brewster, A. S., Michels-Clark, T., Clinger, J. A., Andi, B., Ibrahim, M., Pastor, E., de Lichtenberg, C., Hussein, R., Pollock, C. J., Zhang, M., Stan, C. A., Kroll, T., ... Yano, J. (2017). Drop-on-demand sample delivery for studying biocatalysts in action at X-ray free-electron lasers. *Nature Methods* 2017 14:4, 14(4), 443–449. <https://doi.org/10.1038/nmeth.4195>

He, K., Zhang, X., Ren, S., & Sun, J. (2015). Deep Residual Learning for Image Recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2016-December*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>

Holton, J. M. (2009). A beginner's guide to radiation damage. *Journal of Synchrotron Radiation*, 16(2), 133–142. <https://doi.org/10.1107/S0909049509004361>

Holton, J. M., & Frankel, K. A. (2010). The minimum crystal size needed for a complete diffraction data set. *Acta Crystallographica Section D: Biological Crystallography*, 66(4), 393–408.

Holton, J. M., Classen, S., Frankel, K. A., & Tainer, J. A. (2014). The R-factor gap in macromolecular crystallography: an untapped potential for insights on accurate structures. *The FEBS Journal*, 281(18), 4046–4060.

James, R. W. (1962). *The Optical Principles Of The Diffraction Of X-Rays: The Crystalline State Vol II*. London: Bell & Hyman.

Karplus, P. A., & Diederichs, K. (2012). Linking crystallographic model and data quality. *Science (New York, N.Y.)*, 336(6084), 1030. <https://doi.org/10.1126/SCIENCE.1218231>

Ke, T.-W., Brewster, A. S., Yu, S. X., Ushizima, D., Yang, C., & Sauter, N. K. (2018). A convolutional neural network-based screening tool for X-ray serial crystallography. *Journal of Synchrotron Radiation*, 25(3), 655–670.

Kirkwood, H. J., de Wijn, R., Mills, G., Letrun, R., Kloos, M., Vakili, M., Karnevskiy, M., Ahmed, K., Bean, R. J., Bielecki, J., Dall'Antonia, F., Kim, Y., Kim, C., Koliyadu, J., Round, A., Sato, T., Sikorski, M., Vagovič, P., Sztuk-Dambietz, J., & Mancuso, A. P. (2022). A multi-million image Serial Femtosecond Crystallography dataset collected at the European XFEL. *Scientific Data* 2022 9:1, 9(1), 1–7. <https://doi.org/10.1038/s41597-022-01266-w>

Kmetko, J., Husseini, N. S., Naides, M., Kalinin, Y., & Thorne, R. E. (2006). Quantifying X-ray radiation damage in protein crystals at cryogenic temperatures. *Acta Crystallographica Section D: Biological Crystallography*, 62(9), 1030–1038. <https://doi.org/10.1107/S0907444906023869>

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2323. <https://doi.org/10.1109/5.726791>

Liebschner, D., Afonine, P. v., Baker, M. L., Bunkoczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L. W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. v., ... Adams, P. D. (2019). Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in Phenix. *Urn:Issn:2059-7983*, 75(10), 861–877. <https://doi.org/10.1107/S2059798319011471>

Lieske, J., Cerv, M., Kreida, S., Komadina, D., Fischer, J., Barthelmess, M., Fischer, P., Pakendorf, T., Yefanov, O., Mariani, V., Seine, T., Ross, B. H., Crosas, E., Lorbeer, O., Burkhardt, A., Lane, T. J., Guenther, S., Bergtholdt, J., Schoen, S., ... Meents, A. (2019). On-chip crystallization for serial crystallography experiments and on-chip ligand-binding studies. *IUCrJ*, 6(4), 714–728. <https://doi.org/10.1107/S2052252519007395/MF5034SUP1.PDF>

Leonarski, F., Redford, S., Mozzanica, A., Lopez-Cuenca, C., Panepucci, E., Nass, K., Ozerov, D., Vera, L., Olieric, V., Buntschu, D., & others. (2018). Fast and accurate data collection for macromolecular crystallography using the JUNGFRAU detector. *Nature Methods*, 15(10), 799–804.

Lyubimov, A. Y., Uervirojnangkoorn, M., Zeldin, O. B., Zhou, Q., Zhao, M., Brewster, A. S., Michels-Clark, T., Holton, J. M., Sauter, N. K., Weis, W. I., & Brunger, A. T. (2016). Advances in X-ray free electron laser (XFEL) diffraction data processing applied to the crystal structure of the synaptotagmin-1/SNARE complex. *eLife*, 5(OCTOBER2016). <https://doi.org/10.7554/eLife.18740>

Martiel, I., Müller-Werkmeister, H. M., & Cohen, A. E. (2019). Strategies for sample delivery for femtosecond crystallography. *Acta Crystallographica Section D: Structural Biology*, 75(2), 160–177. <https://doi.org/10.1107/S2059798318017953/BA5296SUP1.PDF>

McPhillips, T. M., McPhillips, S. E., Chiu, H. J., Cohen, A. E., Deacon, A. M., Ellis, P. J., Garman, E., Gonzalez, A., Sauter, N. K., Phizackerley, R. P., Soltis, S. M., & Kuhn, P. (2002). Blu-Ice and the Distributed Control System: software for data acquisition and instrument control at macromolecular crystallography beamlines. *Journal of Synchrotron Radiation*, 9(Pt 6), 401–406. <https://doi.org/10.1107/S0909049502015170>

Milne, C. J., Schietinger, T., Aiba, M., Alarcon, A., Alex, J., Anghel, A., Arsov, V., Beard, C., Beaud, P., Bettoni, S., & others. (2017). SwissFEL: the Swiss X-ray free electron laser. *Applied Sciences*, 7(7), 720.

Nango, E., Kubo, M., Tono, K., & Iwata, S. (2019). Pump-Probe Time-Resolved Serial Femtosecond Crystallography at SACLA: Current Status and Data Collection Strategies. *Applied Sciences 2019, Vol. 9, Page 5505*, 9(24), 5505. <https://doi.org/10.3390/APP9245505>

Park, J., Eom, I., Kang, T. H., Rah, S., Nam, K. H., Park, J., Kim, S., Kwon, S., Park, S. H., Kim, K. S., Hyun, H., Kim, S. N., Lee, E. H., Shin, H., Kim, S., Kim, M. J., Shin, H. J., Ahn, D., Lim, J., ... Kim, S. (2016). Design of a hard X-ray beamline and end-station for pump and probe experiments at Pohang Accelerator Laboratory X-ray Free Electron Laser facility. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, 810, 74–79. <https://doi.org/10.1016/J.NIMA.2015.11.147>

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., Facebook, Z. D., Research, A. I., Lin, Z., Desmaison, A., Antiga, L., Srl, O., & Lerer, A. (2017). *Automatic differentiation in PyTorch*.

Pandey, S., Bean, R., Sato, T., Poudyal, I., Bielecki, J., Cruz Villarreal, J., Yefanov, O., Mariani, V., White, T. A., Kupitz, C., Hunter, M., Abdellatif, M. H., Bajt, S., Bondar, V., Echelmeier, A., Doppler, D., Emons, M., Frank, M., Fromme, R., ... Schmidt, M. (2019). Time-resolved serial femtosecond crystallography at the European XFEL. *Nature Methods 2019 17:1*, 17(1), 73–78. <https://doi.org/10.1038/s41592-019-0628-z>

Pearson, A. R., & Mehrabi, P. (2020). Serial synchrotron crystallography for time-resolved structural biology. *Current Opinion in Structural Biology*, 65, 168–174.
<https://doi.org/10.1016/J.SBI.2020.06.019>

Rahmani, V., Nawaz, S., Pennicard, D., Setty, S. P. R., Graafsma, H., & Boutet, S. (2023). Data reduction for X-ray serial crystallography using machine learning. *Journal of Applied Crystallography*, 56(1), 200–213.
<https://doi.org/10.1107/S1600576722011748/TE5101FIG11THM.GIF>

Raubenheimer, T. O. (2018). The LCLS-II-HE, A High Energy Upgrade of the LCLS-II. *FLS 2018 - Proceedings of the 60th ICFA Advanced Beam Dynamics Workshop on Future Light Sources*, 6–11.
<https://doi.org/10.18429/JACOW-FLS2018-MOP1WA02>

Schmidt, M. (2015). Time-resolved crystallography at X-ray free electron lasers and synchrotron light sources. *Synchrotron Radiation News*, 28(6), 25–30.

Schmidt, S. (2014). GrainSpotter: a fast and robust polycrystalline indexing algorithm.
Urn:Issn:1600-5767, 47(1), 276–284. <https://doi.org/10.1107/S1600576713030185>

Singer, W., Singer, X., Brinkmann, A., Iversen, J., Matheisen, A., Navitski, A., Tamashovich, Y., Michelato, P., & Monaco, L. (2015). Superconducting cavity material for the European XFEL. *Superconductor Science and Technology*, 28(8), 085014. <https://doi.org/10.1088/0953-2048/28/8/085014>

Sauter, N. K., Kern, J., Yano, J., & Holton, J. M. (2020). Towards the spatial resolution of metalloprotein charge states by detailed modeling of XFEL crystallographic diffraction.
Urn:Issn:2059-7983, 76(2), 176–192. <https://doi.org/10.1107/S2059798320000418>

Schulz, E. C., Yorke, B. A., Pearson, A. R., & Mehrabi, P. (2022). Best practices for time-resolved serial synchrotron crystallography. In *Acta crystallographica. Section D, Structural biology* (Vol. 78, Issue 1, pp. 14–29). NLM (Medline). <https://doi.org/10.1107/S2059798321011621>

Šrajer, V., & Schmidt, M. (2017). Watching proteins function with time-resolved x-ray crystallography. In *Journal of Physics D: Applied Physics* (Vol. 50, Issue 37, p. 373001). Institute of Physics Publishing. <https://doi.org/10.1088/1361-6463/aa7d32>

Tanaka, Y., Fukuyama, Y., Šrajer, V., & Schmidt, M. (2017). Watching proteins function with time-resolved x-ray crystallography. *Journal of Physics D: Applied Physics*, 50(37), 373001.
<https://doi.org/10.1088/1361-6463/AA7D32>

Tsai, Y., McPhillips, S. E., González, A., McPhillips, T. M., Zinn, D., Cohen, A. E., Feese, M. D., Bushnell, D., Tiefenbrunn, T., Stout, C. D., Ludaescher, B., Hedman, B., Hodgson, K. O., & Soltis, S. M. (2013). AutoDrug: fully automated macromolecular crystallography workflows for fragment-based drug discovery. *Urn:Issn:0907-4449*, 69(5), 796–803.
<https://doi.org/10.1107/S0907444913001984>

de Wijn, R., Melo, D. V. M., Koua, F. H. M., & Mancuso, A. P. (2022). Potential of Time-Resolved Serial Femtosecond Crystallography Using High Repetition Rate XFEL Sources. *Applied Sciences (Switzerland)*, 12(5), 2551. <https://doi.org/10.3390/app12052551>

Wilson, A. J. C. (1942). Determination of Absolute from Relative X-Ray Intensity Data. *Nature* 1942 150:3796, 150(3796), 152–152. <https://doi.org/10.1038/150152a0>

Winter, G., Waterman, D. G., Parkhurst, J. M., Brewster, A. S., Gildea, R. J., Gerstel, M., Fuentes-Montero, L., Vollmar, M., Michels-Clark, T., Young, I. D., & others. (2018). DIALS: implementation and evaluation of a new integration package. *Acta Crystallographica Section D*, 74(2), 85–97.

Appendix A. Simulation details

A1.1. Crystal models

For each simulated image, a crystal and list of structure factor intensities were modeled using a randomly chosen PBD from the list: 1h74, 1hk5, 1keq, 1ktc, 1lbv, 1nne, 1pdv, 1qtx, 1r03, 1rlk, 1sg8, 1uic, 1uv7, 1vh6, 1xrt, 1yj1, 1yo6, 1z35, 1z6s, 2ar6, 2bh4, 2cc3, 2hu3, 2hyf, 2i8d, 2ibm, 2itu, 2nrz, 2pkg, 2qa4, 2qex, 2qma, 2qt4, 2vj3, 2vuy, 2wox, 2wyf, 2x8i, 2xh6, 2y8k, 2ycf, 2zg2, 2znt, 2zry, 3agy, 3ch7, 3cma, 3cpw, 3dll, 3dxj, 3e6l, 3fj8, 3fl2, 3fyx, 3g8y, 3hfp, 3hxf, 3ilo, 3int, 3k6n, 3l89, 3lke, 3lz7, 3n0w, 3nx, 3oj1, 3t4x, 3tuu, 3u7s, 3uh4, 3uhr, 3vgd, 3woz, 3wpz, 3zbs, 3zg2, 4arq, 4cbc, 4ctn, 4dvn, 4e6i, 4f3x, 4fhm, 4gyk, 4j20, 4m5i, 4m97, 4o09, 4o7s, 4p9h, 4pgu, 4px8, 4qxq, 4rmx, 4wd2, 4xbe, 4xxo, 4ypu, 4z40, 5al4, 5aoo, 5avi, 5dt6, 5g4e, 5g52, 5j77, 5jit, 5o99, 5p9i, 5pj, 5v5k, 5v5v, 5vn7, 5vn9, 5wqg, 5xg2, 6csc.

These PDB files covered the set of space groups P3, P3112, R3:H, P6, P41212, P3221, P321,

P4212, C121, P41, P42212, P3121, C2221, R32:H, P4132, P212121, P6122, I222, P43212, P61, P1, P6522, I212121, P32, P21212, P1211, F432, P213, and I23 and had unit cells ranging in volume from 49,800 Å³ to 48,400,000 Å³. The crystal size was set to 25 μm and the average mosaic domain size was randomly set to 0.05 μm, 0.1 μm, or 0.15 μm. For the resolution training data, the angular mosaic spread of each crystal was randomly chosen along the interval 0.05° to 1°, and for the multi-lattice training data, this range was 0.001° to 0.01°.

A1.2. Detector models

For resolution training datasets, each diffraction pattern was simulated onto either an EIGER 16M or a PILATUS 6M detector format, and the detector distance was randomly sampled along the interval 200 - 300 mm. For the multi-lattice detection training data, a Rayonix 345 format was used, and the detector distance was fixed at 240 mm (according to the experimental geometry it was modeled after). In each simulated detector image, a randomly sized circle of pixels (< 15 mm) was masked to simulate a beam stop. Also, for each image a random selection of 0-5 pixels was chosen, and the pixel values were set to 2¹⁶ photons to simulate “hot” pixels. A second random selection of up to 124 pixels was chosen, and the pixel values were set to 0 to simulate “bad” pixels.

A1.3. Beam model

The total photons per simulated image was 4e11, and each photon had a wavelength of 0.9795Å. The incident beam had a spot size of 30 μm and a divergence of 0.02 mrad.

A1.4. Background scattering

For each simulated image, we computed scattering from 5 mm of air, 25 μm of water, and 25 μm of randomly chosen parasitic source (e.g., glycerol, sucrose, PEG, MDP, DPM, paper, tape, or ice). These background components were summed, and then added to the Bragg scattering, however with a randomly chosen scale factor (between 0.0125 and 1.25) to simulate different experimental conditions and background levels.

A1.5. Simulation timing

Forward scattering simulations were carried out at NERSC using the Perlmutter cluster. Two batch jobs were used to simulate the resolution training data (one per detector model). Each job utilized 64 Perlmutter GPU nodes, each with 4 A100 GPUs, and 16 physical cores per node (4 cores per GPU). In this configuration, the 200,000 PILATUS 6M images and

125,000 EIGER 16M images were simulated in approximately 60 minutes and 90 minutes, respectively.