bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

A UNIFYING BAYESIAN FRAMEWORK FOR MERGING X-RAY DIFFRACTION DATA

Kevin M. Dalton
Department of Molecular & Cellular Biology
Harvard University
Cambridge, MA 02138
kmdalton@fas.harvard.edu

Jack B. Greisman
Department of Molecular & Cellular Biology
Harvard University
Cambridge, MA 02138
greisman@g.harvard.edu

Doeke R. Hekstra
Department of Molecular & Cellular Biology
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
doeke_hekstra@harvard.edu

September 7, 2022

ABSTRACT

Novel X-ray methods are transforming the study of the functional dynamics of biomolecules. Key to this revolution is detection of often subtle conformational changes from diffraction data. Diffraction data contain patterns of bright spots known as reflections. To compute the electron density of a molecule, the intensity of each reflection must be estimated, and redundant observations reduced to consensus intensities. Systematic effects, however, lead to the measurement of equivalent reflections on different scales, corrupting observation of changes in electron density. Here, we present a modern Bayesian solution to this problem, which uses deep learning and variational inference to simultaneously rescale and merge reflection observations. We successfully apply this method to monochromatic and polychromatic single-crystal diffraction data, as well as serial femtosecond crystallography data. We find that this approach is applicable to the analysis of many types of diffraction experiments, while accurately and sensitively detecting subtle dynamics and anomalous scattering.

Keywords X-ray crystallography · Diffraction · Variational inference · Machine learning · Deep learning

X-ray crystallography has revolutionized our understanding of the molecular basis of life by providing atomic-resolution experimental access to the structure and dynamics of macromolecules and their assemblies. In an X-ray diffraction experiment, the electrons of a molecular crystal scatter X-rays, yielding patterns of constructive interference recorded on an X-ray detector. The resulting images contain discrete spots, known as reflections, with intensities proportional to the squared amplitudes of the Fourier components (structure factors) of the crystal's electron density. Each structure factor reports on the electron density at a specific spatial frequency and direction, indexed by triplets of integers termed Miller indices. Estimates of the amplitudes and phases of these structure factors allow one to reconstruct the 3D electron density in the crystal by Fourier synthesis.

Based on these principles, advances in X-ray diffraction now permit direct visualization of macromolecules in action[1] using short X-ray pulses generated at synchrotrons[2, 3] and X-ray Free-Electron Lasers (XFELs)[4, 5]. Full realization of the promise of these methods hinges on the ability to separate signals in X-ray diffraction that result from subtle

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

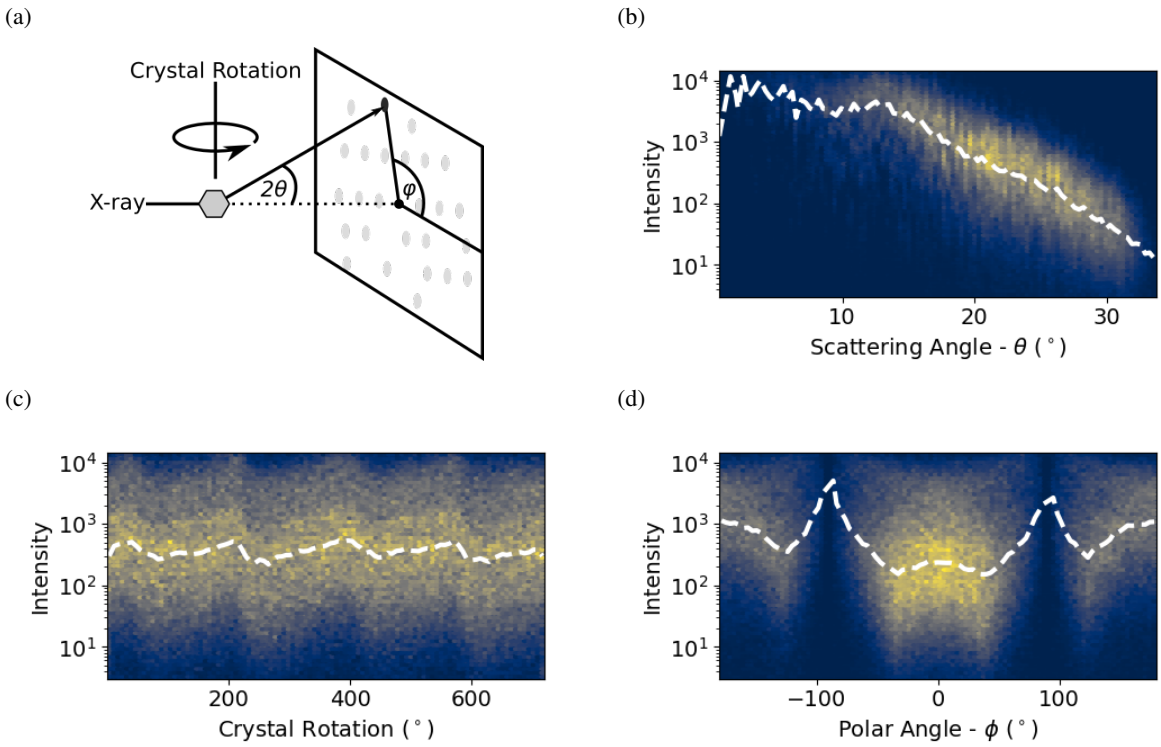


Figure 1: **The intensity of observed reflections depends on physical factors.** (a) Geometry of a conventional diffraction experiment: a crystal (shown as a hexagon) scatters an incident X-ray beam and yields a pattern of reflections (gray spots) on a detector. Three metadata of the measurements are indicated: scattering angle 2θ, crystal rotation angle, and polar angle, ϕ. (b-d) depict the dependence of the observed intensity distributions on these metadata for a hen egg-white lysozyme dataset. The 2-dimensional histograms show the number of counts in each of bins on a logarithmic scale. White dashed lines indicate the median intensity in each x-axis bin. Reflections with $I/\sigma_I <= 0$ were discarded from this analysis. (b) Diffraction intensities decrease with increasing scattering angle, or resolution. (c) Diffraction intensities vary with crystal rotation angle, a proxy for cumulative radiation dose and variations in diffracting volume. (d) Diffraction intensities depend on polar angle due to polarization of the X-ray source and absorption effects.

structural changes from a multitude of systematic errors that can be specific to a crystal, X-ray source, detector, or sample environment[6]. Even under well-controlled experimental conditions, redundant reflections are expressed on the X-ray detector with different scales (Fig. 1). These scales depend non-linearly on the context of each observed reflection as illustrated in Figure 1b-1d. For example, beam properties like intensity fluctuations[7] and polarization[8], crystal imperfections like mosaicity[9] and radiation damage[10], and absorption and scattering of X-rays by material around the crystal all modulate the measured diffraction intensities in a manner which varies throughout the experiment.

Traditionally, these artifacts are accounted for by estimating a series of scale parameters that are intended to explicitly model the physics of the sources of error [11, 6, 12, 13] (see Supplementary Note 3.1 for a description of crystallographic data reduction). The observed intensities are then corrected by each scale parameter to yield scaled intensities. To obtain consensus merged intensities, equivalent observations are then merged by weighted averaging assuming normally distributed errors. This approach thus uses a series of simplifying 'data reduction' steps that work well for standard diffraction experiments at synchrotron beamlines, but are less suited for a rapidly evolving array of next-generation X-ray diffraction experiments.

We took inspiration for an alternative approach from French-Wilson scaling. Although intensities are proportional to squared structure factor amplitudes, negative intensities can result from processing steps such as background subtraction. French-Wilson scaling is commonly used to ensure that inferred structure factor amplitudes are strictly positive [14]. Rather than being based on a physical model of diffraction, this step is based on a Bayesian argument. Namely, structure factor amplitudes are positive and can be expected to follow the so-called Wilson distribution [15]. In a Bayesian sense, the Wilson distribution serves as a prior probability distribution, or prior. This prior can be combined with a statistical

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

model of the true intensity given the observed merged intensity to obtain a posterior probability distribution, or posterior, of the true merged intensity that is strictly positive.

To address the needs of new X-ray diffraction experiments, we introduce a Bayesian model which builds on the paradigm of French and Wilson [14]. We implement this model in an open-source software, Careless, which performs scaling, merging, and French-Wilson corrections in a single step by directly inferring structure factor amplitudes from unscaled, unmerged intensities. In our model, the probability calculation is "forward," predicting integrated intensities from structure factor amplitudes and experimental metadata. As a consequence, analytical tractability of the inference is no longer a concern and the model relating structure factor amplitudes to integrated intensities can be arbitrarily complex and include both explicit physics and machine learning concepts.

We demonstrate that this model can accurately and sensitively extract anomalous signal from single-crystal, monochromatic diffraction at a synchrotron, time-resolved signal from single-crystal, polychromatic diffraction at a synchrotron, and anomalous signal from a serial femtosecond experiment at an XFEL. Our analyses show that this single model can implicitly account for the physical parameters of diffraction experiments with performance competitive with domain-specific, state-of-the-art analysis methods. Although we focus on X-ray diffraction, we believe the same principles can be applied to any diffraction experiment.

Results

Accurate inference of scale parameters and structure factor amplitudes from noisy observations

In a typical diffraction experiment, reflection intensities are recorded along with error estimates and metadata, like crystal orientation, location on the detector, image number, and Miller indices. As shown in Figure 1, the scale on which these observed intensities are expressed can vary systematically due to physical artifacts correlated with the metadata. These different scales must be accounted for in the analysis of diffraction experiments. Here, we present a probabilistic forward model of X-ray diffraction, which we implemented in a software package called Careless. As illustrated in Figure 2a, this model can be abstractly expressed as a so-called probabilistic graphical model. Specifically, the distribution of observable intensities, I_{hi}, for Miller index h in image i, is predicted from structure factor amplitudes F_h and scale factors Σ, which are estimated concurrently. We do so in a Bayesian sense—we estimate the posterior distributions of both the structure factor amplitudes and scale function. We approximate the structure factor amplitudes as statistically independent across Miller indices and use the Wilson distribution as a prior on their magnitudes [15].

By contrast, most contributions to scale factors vary slowly across the data set and are accounted for using a global parametrization. By default, this scale function, Σ, is implemented as a deep neural network which takes the metadata as arguments and predicts the mean and standard deviation of the scale function for each observation (Fig. 2b). To describe measurement error, Careless supports both a normally distributed error model, and a robust Student's t-distributed error model. The implementation of Careless is described in further detail in the Online Methods. The full Bayesian model will typically contain tens of thousands of unique structure factor amplitudes and a dense neural network for the scale function. Use of Markov chain Monte Carlo methods [16], which sample from the posterior, would be computationally prohibitive. Instead, inference is made possible by variational inference [17, 18], in which the parameters of proposed posterior distributions are directly optimized.

We first illustrate the application of the Careless model using a small simulated dataset as shown in Figure 2. As shown in Fig. 2c, we did so by generating true intensities for a toy "crystal" with 3 structure factors of different amplitudes in a mock diffraction experiment with a sharply varying scale function (similar to Fig. 1d). The observed intensities would be recorded with measurement error, yielding a small set of noisy observations (Fig. 2d). Using Careless, we can infer the posterior distributions of the structure factor amplitudes and of the scale factors, and therefore of the true intensities (Fig. 2e). The inferred parameters from Careless show a close correspondence with the true values used to simulate the noisy observations.

Robust inference of anomalous signal from monochromatic diffraction

We next assessed the ability of Careless to extract small crystallographic signals from conventional monochromatic rotation series data (a detailed walk-through of each example is available at https://github.com/Hekstra-Lab/careless-examples). To do so, we applied Careless to a sulfur single-wavelength anomalous diffraction (SAD) data set of lysozyme collected at ambient temperature[19]. It consists of a single 1,440 image rotation series collected in 0.5 degree increments at a low X-ray energy, 6.55 keV, at Advanced Photon Source beamline 24-ID-C. These data are challenging to analyze for several reasons. First, they were collected at a low dose rate to limit radiation damage. Approximately 50% of the pixels on each image did not record a photon. Secondly, there is a shadow from the beam stop mounting

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

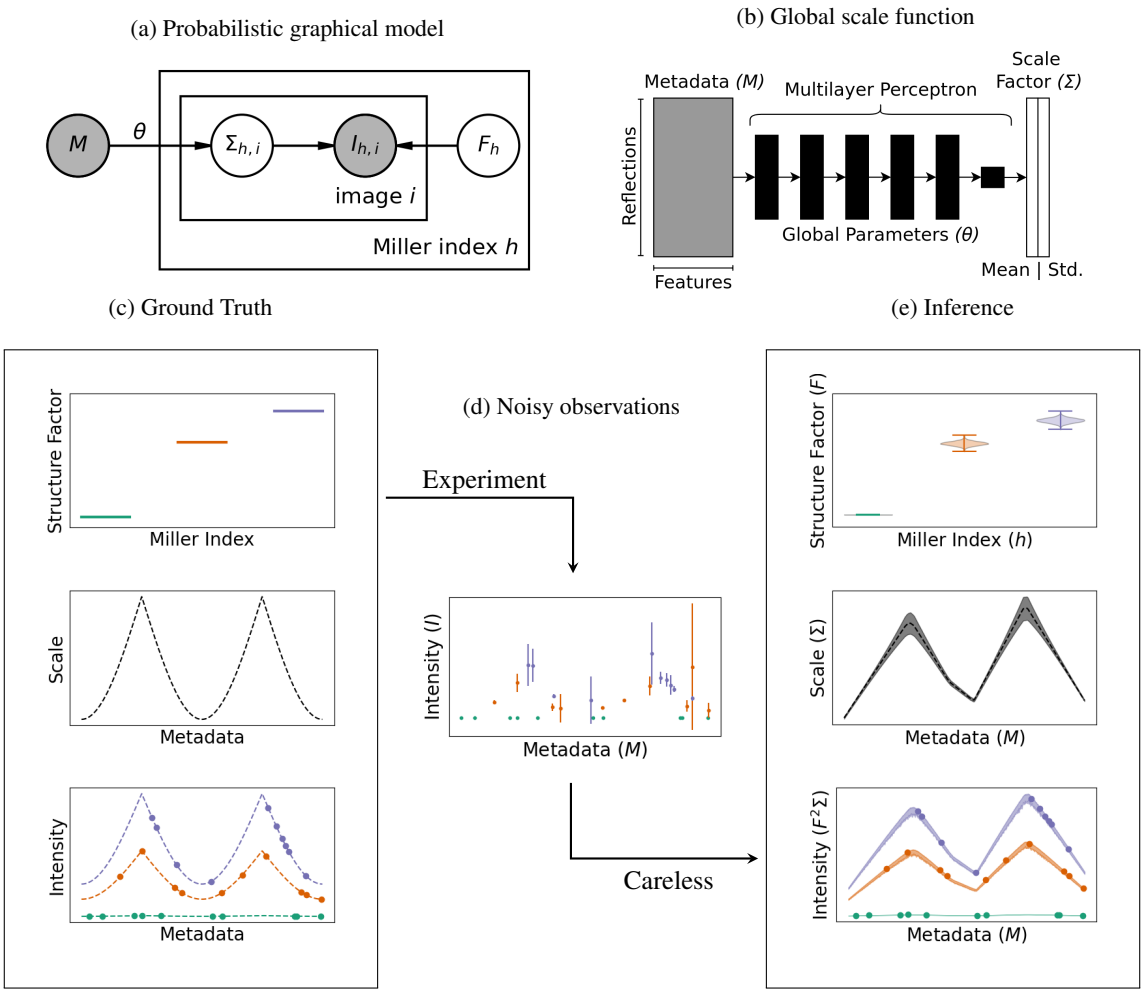


Figure 2: **Estimation of scales and structure factor amplitudes from simulated data.** (a) A probabilistic graphical model summarizes our basic statistical formalism: Careless calculates a probabilistic scale, Σ, as a function of the recorded metadata, M, and learned parameters, θ. Observed intensities, $I_{h,i}$, for Miller index h and image i are modeled as the product of the scale and the square of the structure factor amplitude F_h, that is, as $\Sigma_{hi} \cdot F_h^2$. (b) The global scale function that maps the recorded metadata to the probabilistic scale, Σ, takes the form of a multilayer perceptron. (c-e) Inference of scales and structure factors from simulated data. (c) Input parameters for the simulated data were chosen to recapitulate the non-linear scales observed in diffraction data. (d) Noisy observations were generated from these input parameters, which reflects the measurement errors in a diffraction experiment. (e) This statistical model allows for joint, rather than sequential, inference of the posterior distributions of structure factor amplitudes and scales, and therefore of implied intensity (bottom panel). For this toy example, the inferred values in (e) can be compared to the known ground truth in (c). Error bars in (d) and shaded bands in (e) indicate 95% confidence intervals. Top panel of (e) violin plots of posterior probability with whiskers indicating the extrema of 10,000 samples drawn from the posterior distribution of the inferred F.

bracket on each image near the edge of the detector (in the 2 to 2.2 Å range). Finally, leakage from a higher energy undulator harmonic resulted in a second, smaller diffraction pattern in the center of each image. This artifact, combined with the shadow, means that there are outliers at both high and low resolution in this data set. (Figure 3a).

To address the outliers, we used the cross-validation implemented in Careless (Online Methods 2.7) to select an appropriate degrees of freedom (d.f.) parameter of a Student's *t*-distributed likelihood function for these data. Titrating the number of degrees of freedom, we found that 16 d.f. resulted in the best Spearman correlation coefficient between observations and model predictions (Figure 3b). Accordingly, at 16 degrees of freedom, the structure factor amplitudes exhibit comparatively high half-dataset anomalous correlations (Fig. 3c). As is apparent in Figure 3c, the tuned likelihood function recovers comparable signal to the conventional merging program, Aimless[23].

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

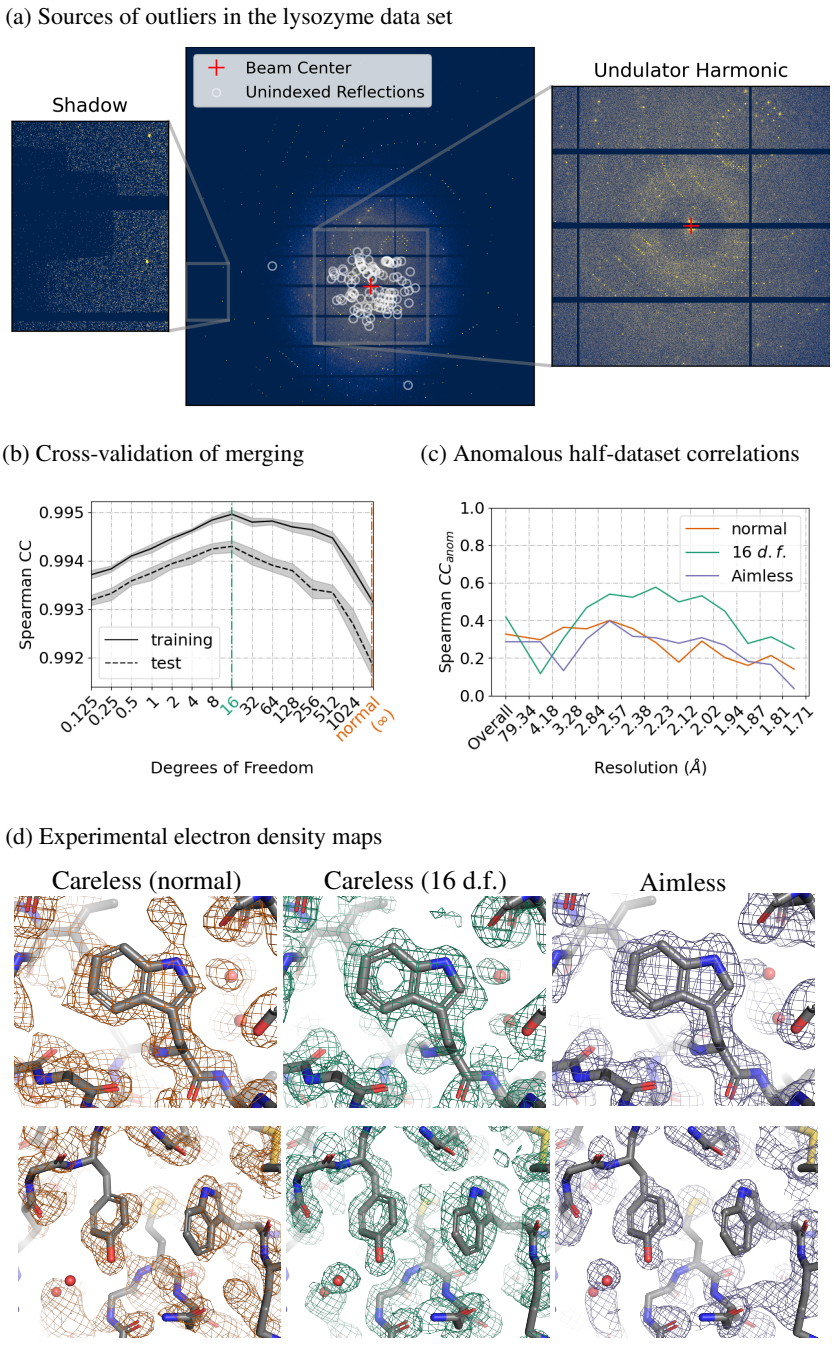


Figure 3: **Accurate processing of sulfur SAD data with Careless**. a) A sample diffraction pattern from the lysozyme data set indicating strong spots which could not be indexed by DIALS [20]. Insets show sources of outliers in the data (left: a beam stop shadow; right: a secondary diffraction pattern resulting from an undulator harmonic). b) Ten-fold cross-validation of merging as a function of the likelihood degrees of freedom. Gray bands: bootstrap 95% confidence intervals from 10 repeats with different randomly chosesn test reflections. c) Spearman correlation coefficient between anomalous differences estimated from half-datasets with jointly trained scale function parameters. (d) Density-modified experimental electron density maps produced with PHENIX Autosol [21] using the sulfur substructure from a reference structure (PDBID: 7L84), contoured at 1.0 σ. Rendered with PyMOL[22].

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

SAD phasing provides a further test of the quality of anomalous signal inferred by Careless. We used Autosol[21] to phase our merging results, comparing the Careless output with a Student's t-distributed likelihood with ∞ or 16 degrees of freedom to the same data merged by the conventional method, Aimless. In order to ensure a consistent origin, we supplied the sulfur atom substructure from the final refined model (PDBID: 7L84) during phasing. Although we provided the heavy atom substructure in the experiments reported here, we were able to phase each of these data sets *ab initio* (Table S1). It is evident from the density-modified experimental maps in Figure 3d that the Careless output with the normally distributed error model (∞ d.f.) is of much lower quality. By contrast, both Careless with 16 d.f. and Aimless produced clearly interpretable experimental maps. The refined heavy atom structure factors yield similar sulfur peak heights (Table S2) for all three processing modalities despite the apparent difference in initial phase estimates for the protein, underscoring the subtle differences in processing quality necessary for successful SAD phasing of this data set. As we illustrate in the online example "Boosting SAD signal with transfer learning", it is possible to further improve the Careless output by using a simple transfer-learning procedure in which the parameters of the scale function are learned by a non-anomalous pre-processing step (see Table S2 for anomalous peak heights). In summary, Careless supports the robust recovery of high-quality experimental electron density maps and anomalous signal in the presence of physical artifacts.

Sensitive detection of time-resolved change from polychromatic diffraction data

Polychromatic (Laue) X-ray diffraction provides an attractive modality for serial and time-resolved X-ray crystallography, as many photons can be delivered in bright femto- or picosecond X-ray pulses [24]. In particular, most reflections in these diffraction snapshots are fully, rather than partially, observed even in still diffraction images. Laue diffraction processing remains, however, a major bottleneck [3] due to its polychromatic nature: The spectrum of a Laue beam is typically peaked with a long tail toward lower energies. This so-called "pink" beam means that reflections recorded at different wavelengths are inherently on different scales. In addition, reflections which lie on the same central ray in reciprocal space will be superposed on the detector. These "harmonic" reflections need to be deconvolved to be merged[25].

Typical polychromatic data reduction software addresses these issues in a series of steps—it uses the experimental geometry to infer which photon energy contributed most strongly to each reflection observation. It then scales the reflections in a wavelength-dependent manner by inferring a wavelength normalization curve related to the spectrum of the X-ray beam [26]. Finally, it deconvolves the contributions to each harmonic reflection by solving a system of linear equations for each image [25]. The need for these steps made it difficult to scale and merge polychromatic data. Not surprisingly, there are no modern open-source merging packages supporting wavelength normalization and harmonic deconvolution.

By contrast, the forward modelling approach implemented in Careless readily extends to the treatment of Laue diffraction. First, to handle wavelength normalization, providing the wavelength of each reflection estimated from experimental geometry as metadata enables the scale function to account for the non-uniform spectrum of the beam. Harmonic deconvolution requires accounting for the fact that the intensity of a reflection is the sum over contributions from all Miller indices lying on the relevant central ray—in the forward probabilistic model implemented in Careless this is a simple extension of the monochromatic case (Online Methods 2.9).

To demonstrate that Careless effectively merges Laue data, we applied it to a time-resolved crystallography data set—20 images from a single crystal of photoactive yellow protein (PYP) in the dark state and 20 images each collected 2 ms after a blue laser pulse. Blue light induces a *trans*- to-*cis* isomerization in the *p*-coumaric acid chromophore in the PYP active site which can be observed in time-resolved experiments[27] (Fig. 4a). We first integrated the Bragg peaks using the commercial Laue data analysis software, Precognition (Renz Research). Then we merged the resulting intensities using Careless.

Careless produces high-quality structure factor amplitudes for this data set, as judged by half-data set correlation coefficients (Fig 4c and 4b). We refined a ground-state model against the 'dark' data starting from a reference model (PDBID: 2PHY) [28], yielding excellent $2F_o - F_c$ electron density maps (Fig. 4d). Using the phases from this refined ground state model, we then constructed unweighted difference maps $|\Delta F_h| \approx (|F_h^{\text{2ms}}| - |F_h^{\text{dark}}|)$, $\phi_h \approx \phi_{h,calc}^{dark}$. As shown in Figure 4e, these maps contain peaks around the PYP chromophore. To better visualize these maps, we applied a previously described weighting procedure [29]. The weighted maps (Figure 4f, 4g) show strong difference density localized to the chromophore, consistent with published models of the dark and excited-state structures[27].

Previous generations of Laue merging software required discarding reflections below a particular I/σ_I cutoff during scaling and merging. Otherwise, the resulting structure factor estimates are not accurate enough to be useful in the analysis of time-resolved structural changes. Here, we applied no such cutoff. Likewise, the appearance of interpretable difference electron density in the absence of a weighting scheme (Fig. 4e) is extraordinary. The ability of Careless to

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

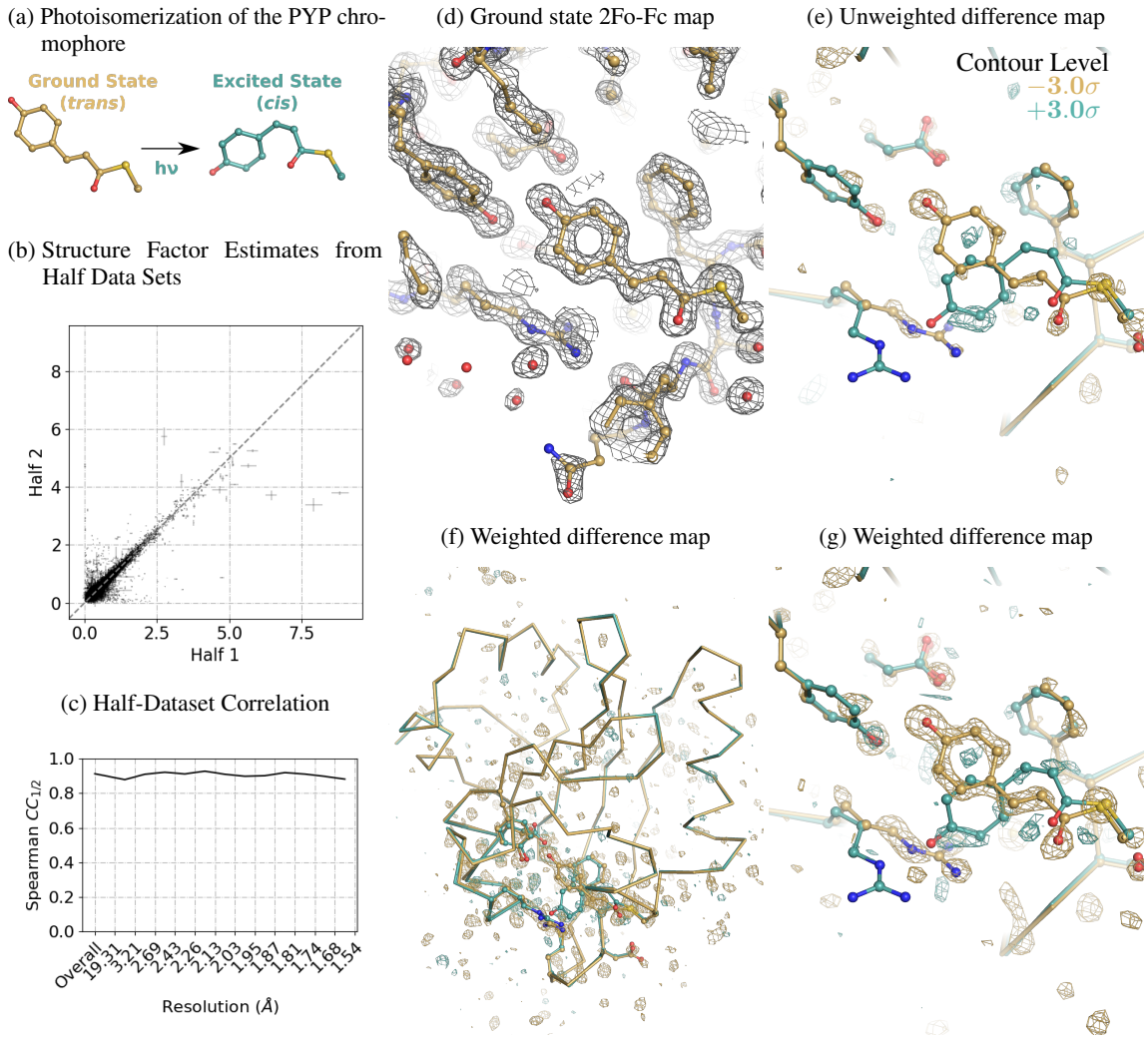


Figure 4: **Careless accurately merges time-resolved polychromatic diffraction data.** (a) When exposed to blue light, the PYP chromophore undergoes *trans*-to-*cis* isomerization. In total, 40 images from a single crystal of PYP were processed: 20 were recorded in the dark state and 20 were recorded 2 ms after illumination with a blue laser pulse. (b) The data were randomly divided in half by image and merged with scale function parameters learned by merging the full data set. Merging with Careless gave excellent correlation between the structure factor estimates of both halves. Error bars indicate the standard deviation of the structure factor posteriors for the half data sets. (c) Half-dataset correlation coefficients as a function of resolution, including both the dark and 2 ms data. (d) Ground-state $2F_o - F_c$ map created by refining the ground-state model (PDBID: 2PHY) against the dark merging results (countoured at 1.5 σ). The phases of this refined model were used for the difference maps in (e-g). (e) $F^{\text{2ms}} - F^{\text{dark}}$ time-resolved difference map showing the accumulation of blue positive density around the excited state chromophore (blue model, PDBID: 3UME) and depletion of the ground state (yellow model, PDBID: 2PHY). (f) $F^{\text{2ms}} - F^{\text{dark}}$ weighted time-resolved difference map showing localization of the difference density to the region surrounding the chromophore. (g) $F^{\text{2ms}} - F^{\text{dark}}$ weighted time-resolved difference map showing large differences around the chromophore. All difference maps are contoured at $\pm 3.0\sigma$.

identify these difference signals demonstrates an unprecedented degree of accuracy and robustness to outliers. As such, Careless improves on the state of the art for the analysis of Laue experiments.

Recovering anomalous signal from a serial experiment at an X-ray Free-Electron Laser

X-ray Free-Electron Lasers (XFELs) are revolutionizing the study of light-driven proteins[30, 31, 32, 33, 34], enzyme microcrystals amenable to rapid mixing[35, 36, 37], and the determination of damage-free structures of difficult-to-

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

crystallize targets [38, 39]. Diffraction data from XFEL sources involve two unique challenges that result from the serial approach commonly used to outrun radiation damage[4]. The first challenge of serial crystallography is that each image originates from a different crystal with a different scattering mass, which diffracts one intense X-ray pulse before structural damage occurs. A completely global scaling model is therefore not appropriate. To overcome this limitation, we exploited the modular design of Careless to incorporate local parameters into the scale function. Specifically, we appended layers with per-image kernel and bias parameters to the global scale function (Fig. 5a). Effectively, these additional layers allow the model to learn a separate scale function for each image. To address the risk of overfitting posed by the additional parameters, we determined the optimal number of image layers by crossvalidation (see Fig. S4 for determination of optimal the number of image-specific layers). A second challenge of serial crystallography is that all images are stills—there is neither time to rotate the crystal during exposure, nor the spectral bandwidth to observe the entirety of each reciprocal lattice point (Fig. 5b). For a still image, the maximal intensity for a given reflection is observed on the detector when the so-called Ewald sphere intersects the reflection centroid. The Ewald offset (EO) measures the degree to which a particular reflection observation deviates from its maximal diffraction condition (the length of the orange arrow in Fig. 5b) and can be estimated from the experimental geometry for each observation. To account for partiality, we hence provided the EO estimates as a metadata.

To test if our model could leverage per-image scale parameters and EO estimates, we applied Careless to unmerged intensities from an XFEL serial crystallography experiment (CXIDB, entry 81). In this experiment[40], a slurry of thermolysin microcrystals was delivered to the XFEL beam by a liquid jet. The data contain significant anomalous signal from the zinc and calcium ions in the structure. We limited analysis to a single run containing 3,160 images. We processed the integrated intensities in Careless using ten global layers and two image layers, both with and without the inclusion of the Ewald offset metadata to evaluate whether its inclusion improves the analysis.

As shown in Figure 5, Careless successfully processes the serial XFEL data. In particular, use of the EO metadatum yielded markedly superior results as judged by the half-dataset correlation coefficient (Fig. 5c) as well as the refinement residuals (Fig. 5d). To verify that including the metadata also improved the information content of the output, we constructed anomalous omit maps after phasing the merged structure factor amplitudes by isomorphous replacement with PDBID 2TLI [41]. Specifically, we omitted the anomalously scattering ions in the reference structure during refinement against the Careless output. The anomalous difference peak heights at the former location of each of the anomalous scatterers, tabulated in Figure 5f, confirm that the inclusion of Ewald offsets not only improved the accuracy of structure factor amplitudes estimates but also yielded anomalous differences on par with XFEL-specific analysis methods.

Discussion

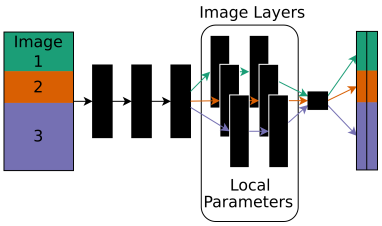
Statistical modeling can account for diverse physical effects

We have shown that Careless successfully scales and merges X-ray diffraction data without the need for explicit physical models of X-ray scattering—comparing favorably to established algorithms tailored to specific crystallography experiments. This begs the question: What sorts of physical effects can the Careless scale function account for using a general-purpose neural network and Wilson's prior distributions on structure factor amplitudes[15]? Though our scale function operates in a high-dimensional space, making it difficult to interrogate directly, we have found that the impact of excluding specific reflection metadata can be used to assess the physical effects that are implicitly accounted for by Careless.

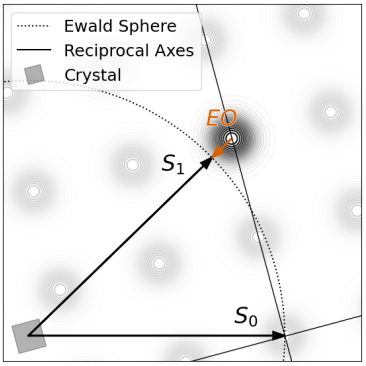
For example, the resolution of each observed reflection is essential for proper data reduction in Careless. This suggests that, among other corrections, the model learns an isotropic scale factor akin to the per-image temperature factors included in most scaling packages[6]. We have also found that in some cases the merging model benefits from the inclusion of the observed Miller indices during scaling, as was the case with the lysozyme and PYP data presented here. This indicates that for some data sets Careless learns an effective anisotropic scaling model. Likewise, the inclusion of the detector positions of the reflections seems to improve merging performance. Since no prior polarization correction has been applied to the PYP and thermolysin intensities used here, this suggests that source polarization is implicitly corrected by the model. Specific crystallographic experiments can also benefit from the inclusion of domain-specific metadata. The success of merging Laue data implies that Careless can learn a function of the spectrum of the X-ray source (Fig. 4). Finally, the XFEL example shows the model is competent to infer partialities in still images (Fig. 5). In principle, the inclusion of local parameters in the form of image layers allows the model to make geometric corrections similar to the partiality models implemented in other packages[42, 43, 44]. The major difference is that our work does not require an explicit model of the line shape of Bragg peaks nor of the resolution-dependence of the peak size (mosaicity).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

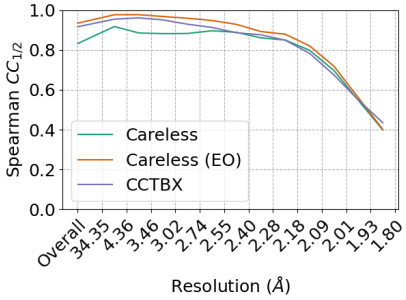
(a) Serial crystallography scaling model



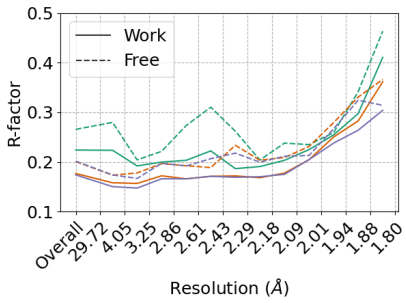
(b) Diffraction condition for stills



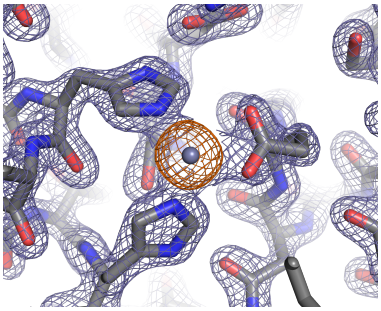
(c) Half data set correlation



(d) Refinement R value



(e) Zn Anomalous Difference Map



(f) Anomalous peak heights

Site	Careless	Careless (EO)	CCTBX
Zn-317	11.58	20.42	20.07
Ca-318	<3.5	5.42	6.17
Ca-319	<3.5	3.61	3.81
Ca-320	4.17	7.32	7.05
Ca-321	<3.5	5.49	7.25

Figure 5: **Careless can leverage geometric metadata to improve XFEL data processing**. (a) By adding image-specific layers to the neural network, Careless can scale diffraction data from serial XFEL experiments. (b) The Ewald offset (EO), shown in red, can optionally be included during processing. S_0 and S_1 represent the directions of the unscattered and diffracted X-rays, respectively. (c) Half-dataset correlation coefficients by resolution bin for data processed in Careless, with and without Ewald offsets, and CCTBX. (d) Refinement R-factors from phenix.refine using Careless, with and without Ewald offsets, and CCTBX. e) Careless $2F_o - F_c$ electron density map, from inclusion of EO, contoured at 2.0 σ (purple mesh) overlayed with thermolysin anomalous omit map contoured at 5.0 σ (orange mesh). f) Peak heights of anomalous scatterers in an anomalous omit map, in σ units, for Careless output with and without Ewald offsets and conventional data processing with CCTBX.

Robust statistics instead of outlier rejection

Occasionally, observed reflections have spurious measured intensities. These outliers can arise from various physical effects during a diffraction experiment, such as ice or salt crystals, detector readout noise, or absorption and scattering by surrounding material[6]. In conventional data reduction, outlier observations are frequently detected and filtered during data processing to improve the estimates of merged intensities and statistics[6, 13]. This step is necessary because the inverse-variance weighting scheme[45] is otherwise easily skewed by spurious observations.

Instead of outlier rejection, Careless employs robust statistical estimators. The processing of the native SAD data from lysozyme (Fig. 3) show its benefit. These data have significant outliers at low and high resolution (Fig. 3a), and the anomalous content is notably improved through the use of a robust error model (Fig. 3c). Importantly, doing so also improves experimental electron density maps compared to the corresponding normally distributed error models (Fig. 3d). The influence of outlier observations can be tuned using the degrees of freedom of the Student's t-distribution during data processing. This approach to handling outlier observations highlights the flexibility of the Careless model and reduces the need for data filtering steps.

Sensitive detection of small structural signals in change-of-state experiments

Our processing of the time-resolved PYP photoisomerization Laue dataset demonstrates that Careless can accurately recover the signal from small structural changes (Fig. 4). The use of Careless in this context involved fitting a common

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

scaling model for both the dark and 2ms datasets, while inferring two separate sets of structure factor amplitudes, one for each state. The quality of the difference maps generated from this processing (Fig. 5e-g) suggests that the common scaling model employed by Careless is effective for analyzing these change-of-state experiments. Importantly, the common scaling model may improve difference maps by ensuring that inferred structure factor amplitudes are on the same scale, producing a balanced difference map with both positive and negative features.

These features suggest that Careless will have strong applications in time-resolved experiments and related change-of-state crystallography experiments. Furthermore, although Careless currently only provides Wilson's distributions as a prior over structure factor amplitudes, it is possible to imagine using stronger priors to further constrain the inference problem. Such priors are an active area of research, and may further improve the sensitivity of Careless to small differences between conditions in time-resolved data sets (see the online example "Using a bivariate prior to exploit correlations between Friedel mates" for a prototype implementation).

Supporting next-generation diffraction experiments

In its current implementation, Careless requires that the full data set reside in memory on a single compute node or accelerator card. This is a significant limitation for free-electron laser applications. The next generation of X-ray free-electron lasers will provide data acquisition rates of $10^3 - 10^6$ diffraction images per second[46] leading to very large data sets with images numbering in the millions or more. In this setting, it would be advantageous to have an online merging algorithm which did not require access to the entire data set at each training step. Currently, the reliance on local parameters to handle serial crystallography data precludes this. However, we are exploring strategies to replace these local parameters with a global function. This will enable Careless to be implemented with stochastic training[47], in which gradient descent is conducted on a subset of the data at each iteration. This training paradigm allows variational inference to be used for data sets too large to fit in memory. With stochastic training, Careless will be an excellent candidate for merging large XFEL data sets on-the-fly during data acquisition.

In summary, we have described a general, extensible framework for inference of structure factor amplitudes from integrated X-ray diffraction intensities. We find the approach to be accurate, and applicable to a wide range of X-ray diffraction modalities. Careless is modular and open-source. We encourage users to report their experiences with downstream software and to contribute extensions through github.com/Hekstra-Lab/careless. Careless provides a foundation for the ongoing development and systematic application of advanced probabilistic models to the analysis of ever more powerful diffraction experiments.

Acknowledgements

We acknowledge the open source software projects used in this work including Daft[48], DIALS[20], GEMMI[49], Matplotlib[50], Numpy[51], Pandas[52], reciprocalspaceship[53], SciPy[54], Seaborn[55], TensorFlow[56], TensorFlow Probability[57] . We thank Vukica Srajer and Marius Schmidt for providing Laue time-resolved data, Nick Sauter and Aaron Brewster for helpful discussions, and T.J. Lane and Takanori Nakane for comments on the manuscript. We thank Derek Mendez for advice on CCTBX. We thank the staff at the Northeastern Collaborative Access Team (NE-CAT), beamline 24-ID-C of the Advanced Photon Source for assistance with room-temperature crystallography, in particular Igor Kourinov. NE-CAT beamlines are supported by the National Institute of General Medical Sciences, NIH (P30 GM124165), using resources of the Advanced Photon Source, a U.S. Department of Energy (DOE) Office of Science User Facility operated for the DOE Office of Science by Argonne National Laboratory under Contract No. DE-AC02-06CH11357. This work was supported by the Searle Scholarship Program (SSP-2018-3240) and a Merck Fellowship (338034) from the George W. Merck Fund of the New York Community Trust (both to D.R.H.). J.B.G. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303. K.M.D. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

Data Availability

Careless is available from our GitHub page (`https://github.com/Hekstra-Lab/careless`). The refined lysozyme structure is deposited in the Protein Data Bank (PDBID: 7L84), and the raw lysozyme diffraction images are available through the SBGrid Data Bank (ID: 816). The three data sets discussed in this manuscript have been adapted into examples available through the `careless-examples` GitHub page (`https://github.com/Hekstra-Lab/careless`) including the unmerged diffraction data and relevant merging scripts. The source code and intermediate analysis used to generate all figures and tables is freely available from Zenodo (`https://doi.org/10.5281/zenodo.6408750`).

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Code Availability

The algorithm described herein is implemented in a python package which is available from our GitHub page (`https://github.com/Hekstra-Lab/careless`). It can be installed on Mac OS or Linux with the popular python package manager, `pip`.

Author Contributions

K.M.D. and D.R.H. conceived the project. K.M.D. conceived and implemented the inference algorithm. J.B.G. contributed to software development. K.M.D. drafted the manuscript. All of the authors contributed to analysis and edited the paper.

References

[1] Vukica Šrajer and Marius Schmidt. Watching proteins function with time-resolved x-ray crystallography. *Journal of physics D: Applied physics*, 50(37):373001, 09 2017.

[2] T. Graber, S. Anderson, H. Brewer, Y.-S. Chen, H. S. Cho, N. Dashdorj, R. W. Henning, I. Kosheleva, G. Macha, M. Meron, R. Pahl, Z. Ren, S. Ruan, F. Schotte, V. Šrajer, P. J. Viccaro, F. Westferro, P. Anfinrud, and K. Moffat. BioCARS: a synchrotron resource for time-resolved X-ray science. *Journal of Synchrotron Radiation*, 18(4):658–670, Jul 2011.

[3] Keith Moffat. Laue diffraction and time-resolved crystallography: a personal history. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 377(2147):20180243, 2019.

[4] Henry N. Chapman, Petra Fromme, Anton Barty, Thomas A. White, Richard A. Kirian, Andrew Aquila, Mark S. Hunter, Joachim Schulz, Daniel P. DePonte, Uwe Weierstall, R. Bruce Doak, Filipe R. N. C. Maia, Andrew V. Martin, Ilme Schlichting, Lukas Lomb, Nicola Coppola, Robert L. Shoeman, Sascha W. Epp, Robert Hartmann, Daniel Rolles, Artem Rudenko, Lutz Foucar, Nils Kimmel, Georg Weidenspointner, Peter Holl, Mengning Liang, Miriam Barthelmess, Carl Caleman, Sébastien Boutet, Michael J. Bogan, Jacek Krzywinski, Christoph Bostedt, Saša Bajt, Lars Gumprecht, Benedikt Rudek, Benjamin Erk, Carlo Schmidt, André Hömke, Christian Reich, Daniel Pietschner, Lothar Strüder, Günter Hauser, Hubert Gorke, Joachim Ullrich, Sven Herrmann, Gerhard Schaller, Florian Schopper, Heike Soltau, Kai-Uwe Kühnel, Marc Messerschmidt, John D. Bozek, Stefan P. Hau-Riege, Matthias Frank, Christina Y. Hampton, Raymond G. Sierra, Dmitri Starodub, Garth J. Williams, Janos Hajdu, Nicusor Timneanu, M. Marvin Seibert, Jakob Andreasson, Andrea Rocker, Olof Jönsson, Martin Svenda, Stephan Stern, Karol Nass, Robert Andritschke, Claus-Dieter Schröter, Faton Krasniqi, Mario Bott, Kevin E. Schmidt, Xiaoyu Wang, Ingo Grotjohann, James M. Holton, Thomas R. M. Barends, Richard Neutze, Stefano Marchesini, Raimund Fromme, Sebastian Schorb, Daniela Rupp, Marcus Adolph, Tais Gorkhover, Inger Andersson, Helmut Hirsemann, Guillaume Potdevin, Heinz Graafsma, Björn Nilsson, and John C. H. Spence. Femtosecond x-ray protein nanocrystallography. *Nature*, 470(7332):73–77, 2011.

[5] Suraj Pandey, Richard Bean, Tokushi Sato, Ishwor Poudyal, Johan Bielecki, Jorvani Cruz Villarreal, Oleksandr Yefanov, Valerio Mariani, Thomas A. White, Christopher Kupitz, Mark Hunter, Mohamed H. Abdellatif, Saša Bajt, Valerii Bondar, Austin Echelmeier, Diandra Doppler, Moritz Emons, Matthias Frank, Raimund Fromme, Yaroslav Gevorkov, Gabriele Giovanetti, Man Jiang, Daihyun Kim, Yoonhee Kim, Henry Kirkwood, Anna Klimovskaia, Juraj Knoska, Faisal H. M. Koua, Romain Letrun, Stella Lisova, Luis Maia, Victoria Mazalova, Domingo Meza, Thomas Michelat, Abbas Ourmazd, Guido Palmer, Marco Ramilli, Robin Schubert, Peter Schwander, Alessandro Silenzi, Jolanta Sztuk-Dambietz, Alexandra Tolstikova, Henry N. Chapman, Alexandra Ros, Anton Barty, Petra Fromme, Adrian P. Mancuso, and Marius Schmidt. Time-resolved serial femtosecond crystallography at the european xfel. *Nature Methods*, 17(1):73–78, 2020.

[6] Philip Evans. Scaling and assessment of data quality. *Acta Crystallographica Section D*, 62(1):72–82, Jan 2006.

[7] R. Bonifacio, L. De Salvo, P. Pierini, N. Piovella, and C. Pellegrini. Spectrum, temporal structure, and fluctuations in a high-gain free-electron laser starting from noise. *Phys. Rev. Lett.*, 73:70–73, Jul 1994.

[8] L. V. Azároff. Polarization correction for crystal-monochromatized X-radiation. *Acta Crystallographica*, 8(11):701–704, Nov 1955.

[9] Colin Nave. A Description of Imperfections in Protein Crystals. *Acta Crystallographica Section D*, 54(5):848–853, Sep 1998.

[10] Elspeth F. Garman. Radiation damage in macromolecular crystallography: what is it and why should we care? *Acta Crystallographica Section D*, 66(4):339–351, Apr 2010.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[11] Zbyszek Otwinowski, Dominika Borek, Wladyslaw Majewski, and Wladek Minor. Multiparametric scaling of diffraction intensities. *Acta Crystallographica Section A*, 59(3):228–234, May 2003.

[12] Wolfgang Kabsch. Integration, scaling, space-group assignment and post-refinement. *Acta Crystallographica Section D*, 66(2):133–144, Feb 2010.

[13] James Beilsten-Edmands, Graeme Winter, Richard Gildea, James Parkhurst, David Waterman, and Gwyndaf Evans. Scaling diffraction data in the *DIALS* software package: algorithms and new approaches for multi-crystal scaling. *Acta Crystallographica Section D*, 76(4):385–399, Apr 2020.

[14] S. French and K. Wilson. On the treatment of negative intensity observations. *Acta Crystallographica Section A*, 34(4):517–525, 1978.

[15] A. J. C. Wilson. The probability distribution of x-ray intensities. *Acta Crystallographica*, 2(5):318–321, 1949.

[16] Charles J. Geyer. Practical Markov Chain Monte Carlo. *Statistical Science*, 7(4):473 – 483, 1992.

[17] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, November 1999.

[18] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. arXiv: 1601.00670.

[19] Jack B. Greisman, Kevin M. Dalton, Candice J. Sheehan, Margaret A. Klureza, and Doeke R. Hekstra. Native SAD phasing at room temperature. *bioRxiv*, page 2021.12.13.472485.

[20] G. Winter, D. G. Waterman, J. M. Parkhurst, A. S. Brewster, R. J. Gildea, M. Gerstel, L. Fuentes-Montero, M. Vollmar, T. Michels-Clark, I. D. Young, N. K. Sauter, and G. Evans. DIALS: implementation and evaluation of a new integration package. 74(2):85–97. Number: 2 Publisher: International Union of Crystallography.

[21] T. C. Terwilliger, P. D. Adams, R. J. Read, A. J. McCoy, N. W. Moriarty, R. W. Grosse-Kunstleve, P. V. Afonine, P. H. Zwart, and L.-W. Hung. Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallographica Section D: Biological Crystallography*, 65(6):582–601, June 2009.

[22] Schrödinger, LLC. The PyMOL molecular graphics system, version 1.8. November 2015.

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

[24] A. Meents, M. O. Wiedorn, V. Srajer, R. Henning, I. Sarrou, J. Bergtholdt, M. Barthelmess, P. Y. A. Reinke, D. Dierksmeyer, A. Tolstikova, S. Schaible, M. Messerschmidt, C. M. Ogata, D. J. Kissick, M. H. Taft, D. J. Manstein, J. Lieske, D. Oberthuer, R. F. Fischetti, and H. N. Chapman. Pink-beam serial crystallography. *Nature Communications*, 8(1):1281, 2017.

[25] Z. Ren and K. Moffat. Deconvolution of energy overlaps in laue diffraction. *Journal of Applied Crystallography*, 28(5):482–494, 1995.

[26] Z. Ren, D. Bourgeois, J. R. Helliwell, K. Moffat, V. Srajer, and B. L. Stoddard. Laue crystallography: coming of age. *Journal of Synchrotron Radiation*, 6(4):891–917, 1999.

[27] Ulrich K. Genick, Gloria E. O. Borgstahl, Kingman Ng, Zhong Ren, Claude Pradervand, Patrick M. Burke, Vukica Šrajer, Tsu-Yi Teng, Wilfried Schildkamp, Duncan E. McRee, Keith Moffat, and Elizabeth D. Getzoff. Structure of a protein photocycle intermediate by millisecond time-resolved crystallography. *Science*, 275(5305):1471–1475, 1997.

[28] Gloria E. O. Borgstahl, DeWight R. Williams, and Elizabeth D. Getzoff. 1.4 Å structure of photoactive yellow protein, a cytosolic photoreceptor: Unusual fold, active site, and chromophore. *Biochemistry*, 34(19):6278–6287, 1995.

[29] Doeke R. Hekstra, K. Ian White, Michael A. Socolich, Robert W. Henning, Vukica Šrajer, and Rama Ranganathan. Electric-field-stimulated protein mechanics. *Nature*, 540(7633):400–405, 2016.

[30] Jason Tenboer, Shibom Basu, Nadia Zatsepin, Kanupriya Pande, Despina Milathianaki, Matthias Frank, Mark Hunter, Sébastien Boutet, Garth J Williams, Jason E Koglin, Dominik Oberthuer, Michael Heymann, Christopher Kupitz, Chelsie Conrad, Jesse Coe, Shatabdi Roy-Chowdhury, Uwe Weierstall, Daniel James, Dingjie Wang, Thomas Grant, Anton Barty, Oleksandr Yefanov, Jennifer Scales, Cornelius Gati, Carolin Seuring, Vukica Srajer, Robert Henning, Peter Schwander, Raimund Fromme, Abbas Ourmazd, Keith Moffat, Jasper J Van Thor, John C H Spence, Petra Fromme, Henry N Chapman, and Marius Schmidt. Time-resolved serial crystallography captures high-resolution intermediates of photoactive yellow protein. *Science*, 346(6214):1242–1246, 12 2014.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[31] Kanupriya Pande, Christopher D M Hutchison, Gerrit Groenhof, Andy Aquila, Josef S Robinson, Jason Tenboer, Shibom Basu, Sébastien Boutet, Daniel P DePonte, Mengning Liang, Thomas A White, Nadia A Zatsepin, Oleksandr Yefanov, Dmitry Morozov, Dominik Oberthuer, Cornelius Gati, Ganesh Subramanian, Daniel James, Yun Zhao, Jake Koralek, Jennifer Brayshaw, Christopher Kupitz, Chelsie Conrad, Shatabdi Roy-Chowdhury, Jesse D Coe, Markus Metz, Paulraj Lourdu Xavier, Thomas D Grant, Jason E Koglin, Gihan Ketawala, Raimund Fromme, Vukica Šrajer, Robert Henning, John C H Spence, Abbas Ourmazd, Peter Schwander, Uwe Weierstall, Matthias Frank, Petra Fromme, Anton Barty, Henry N Chapman, Keith Moffat, Jasper J van Thor, and Marius Schmidt. Femtosecond structural dynamics drives the trans/cis isomerization in photoactive yellow protein. *Science*, 352(6286):725–729, 05 2016.

[32] Eriko Nango, Antoine Royant, Minoru Kubo, Takanori Nakane, Cecilia Wickstrand, Tetsunari Kimura, Tomoyuki Tanaka, Kensuke Tono, Changyong Song, Rie Tanaka, Toshi Arima, Ayumi Yamashita, Jun Kobayashi, Toshiaki Hosaka, Eiichi Mizohata, Przemyslaw Nogly, Michihiro Sugahara, Daewoong Nam, Takashi Nomura, Tatsuro Shimamura, Dohyun Im, Takaaki Fujiwara, Yasuaki Yamanaka, Byeonghyun Jeon, Tomohiro Nishizawa, Kazumasa Oda, Masahiro Fukuda, Rebecka Andersson, Petra Båth, Robert Dods, Jan Davidsson, Shigeru Matsuoka, Satoshi Kawatake, Michio Murata, Osamu Nureki, Shigeki Owada, Takashi Kameshima, Takaki Hatsui, Yasumasa Joti, Gebhard Schertler, Makina Yabashi, Ana-Nicoleta Bondar, Jörg Standfuss, Richard Neutze, and So Iwata. A three-dimensional movie of structural changes in bacteriorhodopsin. *Science*, 354(6319):1552–1557, Dec 2016.

[33] Michihiro Suga, Fusamichi Akita, Michihiro Sugahara, Minoru Kubo, Yoshiki Nakajima, Takanori Nakane, Keitaro Yamashita, Yasufumi Umena, Makoto Nakabayashi, Takahiro Yamane, Takamitsu Nakano, Mamoru Suzuki, Tetsuya Masuda, Shigeyuki Inoue, Tetsunari Kimura, Takashi Nomura, Shinichiro Yonekura, Long-Jiang Yu, Tomohiro Sakamoto, Taiki Motomura, Jing-Hua Chen, Yuki Kato, Takumi Noguchi, Kensuke Tono, Yasumasa Joti, Takashi Kameshima, Takaki Hatsui, Eriko Nango, Rie Tanaka, Hisashi Naitow, Yoshinori Matsuura, Ayumi Yamashita, Masaki Yamamoto, Osamu Nureki, Makina Yabashi, Tetsuya Ishikawa, So Iwata, and Jian-Ren Shen. Light-induced structural changes and the site of o=o bond formation in psii caught by xfel. *Nature*, 543(7643):131–135, 2017.

[34] Atsuhiro Shimada, Minoru Kubo, Seiki Baba, Keitaro Yamashita, Kunio Hirata, Go Ueno, Takashi Nomura, Tetsunari Kimura, Kyoko Shinzawa-Itoh, Junpei Baba, Keita Hatano, Yuki Eto, Akari Miyamoto, Hironori Murakami, Takashi Kumasaka, Shigeki Owada, Kensuke Tono, Makina Yabashi, Yoshihiro Yamaguchi, Sachiko Yanagisawa, Miyuki Sakaguchi, Takashi Ogura, Ryo Komiya, Jiwang Yan, Eiki Yamashita, Masaki Yamamoto, Hideo Ago, Shinya Yoshikawa, and Tomitake Tsukihara. A nanosecond time-resolved xfel analysis of structural changes associated with co release from cytochrome c oxidase. *Science advances*, 3(7):e1603042–e1603042, 07 2017.

[35] J. R. Stagno, Y. Liu, Y. R. Bhandari, C. E. Conrad, S. Panja, M. Swain, L. Fan, G. Nelson, C. Li, D. R. Wendel, T. A. White, J. D. Coe, M. O. Wiedorn, J. Knoska, D. Oberthuer, R. A. Tuckey, P. Yu, M. Dyba, S. G. Tarasov, U. Weierstall, T. D. Grant, C. D. Schwieters, J. Zhang, A. R. Ferré-D'Amaré, P. Fromme, D. E. Draper, M. Liang, M. S. Hunter, S. Boutet, K. Tan, X. Zuo, X. Ji, A. Barty, N. A. Zatsepin, H. N. Chapman, J. C. H. Spence, S. A. Woodson, and Y. X. Wang. Structures of riboswitch rna reaction states by mix-and-inject xfel serial crystallography. *Nature*, 541(7636):242–246, 2017.

[36] Jose L. Olmos, Suraj Pandey, Jose M. Martin-Garcia, George Calvey, Andrea Katz, Juraj Knoska, Christopher Kupitz, Mark S. Hunter, Mengning Liang, Dominik Oberthuer, Oleksandr Yefanov, Max Wiedorn, Michael Heyman, Mark Holl, Kanupriya Pande, Anton Barty, Mitchell D. Miller, Stephan Stern, Shatabdi Roy-Chowdhury, Jesse Coe, Nirupa Nagaratnam, James Zook, Jacob Verburgt, Tyler Norwood, Ishwor Poudyal, David Xu, Jason Koglin, Matthew H. Seaberg, Yun Zhao, Saša Bajt, Thomas Grant, Valerio Mariani, Garrett Nelson, Ganesh Subramanian, Euiyoung Bae, Raimund Fromme, Russell Fung, Peter Schwander, Matthias Frank, Thomas A. White, Uwe Weierstall, Nadia Zatsepin, John Spence, Petra Fromme, Henry N. Chapman, Lois Pollack, Lee Tremblay, Abbas Ourmazd, George N. Phillips, and Marius Schmidt. Enzyme intermediates captured "on the fly"by mix-and-inject serial crystallography. *BMC Biology*, 16(1):59, 2018.

[37] Medhanjali Dasgupta, Dominik Budday, Saulo H. P. de Oliveira, Peter Madzelan, Darya Marchany-Rivera, Javier Seravalli, Brandon Hayes, Raymond G. Sierra, Sébastien Boutet, Mark S. Hunter, Roberto Alonso-Mori, Alexander Batyuk, Jennifer Wierman, Artem Lyubimov, Aaron S. Brewster, Nicholas K. Sauter, Gregory A. Applegate, Virendra K. Tiwari, David B. Berkowitz, Michael C. Thompson, Aina E. Cohen, James S. Fraser, Michael E. Wall, Henry van den Bedem, and Mark A. Wilson. Mix-and-inject xfel crystallography reveals gated conformational dynamics during enzyme catalysis. *Proceedings of the National Academy of Sciences*, 116(51):25634–25640, 2019.

[38] Yanyong Kang, X Edward Zhou, Xiang Gao, Yuanzheng He, Wei Liu, Andrii Ishchenko, Anton Barty, Thomas A White, Oleksandr Yefanov, Gye Won Han, Qingping Xu, Parker W de Waal, Jiyuan Ke, M H Eileen Tan, Chenghai Zhang, Arne Moeller, Graham M West, Bruce D Pascal, Ned Van Eps, Lydia N Caro, Sergey A Vishnivetskiy,

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Regina J Lee, Kelly M Suino-Powell, Xin Gu, Kuntal Pal, Jinming Ma, Xiaoyong Zhi, Sébastien Boutet, Garth J Williams, Marc Messerschmidt, Cornelius Gati, Nadia A Zatsepin, Dingjie Wang, Daniel James, Shibom Basu, Shatabdi Roy-Chowdhury, Chelsie E Conrad, Jesse Coe, Haiguang Liu, Stella Lisova, Christopher Kupitz, Ingo Grotjohann, Raimund Fromme, Yi Jiang, Minjia Tan, Huaiyu Yang, Jun Li, Meitian Wang, Zhong Zheng, Dianfan Li, Nicole Howe, Yingming Zhao, Jörg Standfuss, Kay Diederichs, Yuhui Dong, Clinton S Potter, Bridget Carragher, Martin Caffrey, Hualiang Jiang, Henry N Chapman, John C H Spence, Petra Fromme, Uwe Weierstall, Oliver P Ernst, Vsevolod Katritch, Vsevolod V Gurevich, Patrick R Griffin, Wayne L Hubbell, Raymond C Stevens, Vadim Cherezov, Karsten Melcher, and H Eric Xu. Crystal structure of rhodopsin bound to arrestin by femtosecond x-ray laser. *Nature*, 523(7562):561–567, 07 2015.

[39] Alexander Batyuk, Lorenzo Galli, Andrii Ishchenko, Gye Won Han, Cornelius Gati, Petr A Popov, Ming-Yue Lee, Benjamin Stauch, Thomas A White, Anton Barty, Andrew Aquila, Mark S Hunter, Mengning Liang, Sébastien Boutet, Mengchen Pu, Zhi-Jie Liu, Garrett Nelson, Daniel James, Chufeng Li, Yun Zhao, John C H Spence, Wei Liu, Petra Fromme, Vsevolod Katritch, Uwe Weierstall, Raymond C Stevens, and Vadim Cherezov. Native phasing of x-ray free-electron laser data for a g protein-coupled receptor. *Science advances*, 2(9):e1600292–e1600292, 09 2016.

[40] Jan Kern, Rosalie Tran, Roberto Alonso-Mori, Sergey Koroidov, Nathaniel Echols, Johan Hattne, Mohamed Ibrahim, Sheraz Gul, Hartawan Laksmono, Raymond G. Sierra, Richard J. Gildea, Guangye Han, Julia Hellmich, Benedikt Lassalle-Kaiser, Ruchira Chatterjee, Aaron S. Brewster, Claudiu A. Stan, Carina Glöckner, Alyssa Lampe, Dörte DiFiore, Despina Milathianaki, Alan R. Fry, M. Marvin Seibert, Jason E. Koglin, Erik Gallo, Jens Uhlig, Dimosthenis Sokaras, Tsu-Chien Weng, Petrus H. Zwart, David E. Skinner, Michael J. Bogan, Marc Messerschmidt, Pieter Glatzel, Garth J. Williams, Sébastien Boutet, Paul D. Adams, Athina Zouni, Johannes Messinger, Nicholas K. Sauter, Uwe Bergmann, Junko Yano, and Vittal K. Yachandra. Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nature Communications*, 5(1):4371, July 2014.

[41] Andrew C. English, Sarah H. Done, Leo S. D. Caves, Colin R. Groom, and Roderick E. Hubbard. Locating interaction sites on proteins: The crystal structure of thermolysin soaked in 2% to 100% isopropanol. *Proteins: Structure, Function, and Bioinformatics*, 37(4):628–640, 1999.

[42] Monarin Uervirojnangkoorn, Oliver B Zeldin, Artem Y Lyubimov, Johan Hattne, Aaron S Brewster, Nicholas K Sauter, Axel T Brunger, and William I Weis. Enabling X-ray free electron laser crystallography for challenging biological systems from a limited number of crystals. *eLife*, 4:e05421, March 2015.

[43] Nicholas K. Sauter. XFEL diffraction: developing processing methods to optimize data quality. *Journal of Synchrotron Radiation*, 22(2):239–248, Mar 2015.

[44] Thomas A. White. Post-refinement method for snapshot serial crystallography. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1647):20130330, July 2014.

[45] W. C. Hamilton, J. S. Rollett, and R. A. Sparks. On the relative scaling of X-ray photographs. *Acta Crystallographica*, 18(1):129–130, Jan 1965.

[46] Max O. Wiedorn, Dominik Oberthür, Richard Bean, Robin Schubert, Nadine Werner, Brian Abbey, Martin Aepfelbacher, Luigi Adriano, Aschkan Allahgholi, Nasser Al-Qudami, Jakob Andreasson, Steve Aplin, Salah Awel, Kartik Ayyer, Saša Bajt, Imrich Barák, Sadia Bari, Johan Bielecki, Sabine Botha, Djelloul Boukhelef, Wolfgang Brehm, Sandor Brockhauser, Igor Cheviakov, Matthew A. Coleman, Francisco Cruz-Mazo, Cyril Danilevski, Connie Darmanin, R. Bruce Doak, Martin Domaracky, Katerina Dörner, Yang Du, Hans Fangohr, Holger Fleckenstein, Matthias Frank, Petra Fromme, Alfonso M. Gañán-Calvo, Yaroslav Gevorkov, Klaus Giewekemeyer, Helen Mary Ginn, Heinz Graafsma, Rita Graceffa, Dominic Greiffenberg, Lars Gumprecht, Peter Göttlicher, Janos Hajdu, Steffen Hauf, Michael Heymann, Susannah Holmes, Daniel A. Horke, Mark S. Hunter, Siegfried Imlau, Alexander Kaukher, Yoonhee Kim, Alexander Klyuev, Juraj Knoška, Bostjan Kobe, Manuela Kuhn, Christopher Kupitz, Jochen Küpper, Janine Mia Lahey-Rudolph, Torsten Laurus, Karoline Le Cong, Romain Letrun, P. Lourdu Xavier, Luis Maia, Filipe R. N. C. Maia, Valerio Mariani, Marc Messerschmidt, Markus Metz, Davide Mezza, Thomas Michelat, Grant Mills, Diana C. F. Monteiro, Andrew Morgan, Kerstin Mühlig, Anna Munke, Astrid Münnich, Julia Nette, Keith A. Nugent, Theresa Nuguid, Allen M. Orville, Suraj Pandey, Gisel Pena, Pablo Villanueva-Perez, Jennifer Poehlsen, Gianpietro Previtali, Lars Redecke, Winnie Maria Riekehr, Holger Rohde, Adam Round, Tatiana Safenreiter, Iosifina Sarrou, Tokushi Sato, Marius Schmidt, Bernd Schmitt, Robert Schönherr, Joachim Schulz, Jonas A. Sellberg, M. Marvin Seibert, Carolin Seuring, Megan L. Shelby, Robert L. Shoeman, Marcin Sikorski, Alessandro Silenzi, Claudiu A. Stan, Xintian Shi, Stephan Stern, Jola Sztuk-Dambietz, Janusz Szuba, Aleksandra Tolstikova, Martin Trebbin, Ulrich Trunk, Patrik Vagovic, Thomas Ve, Britta Weinhausen, Thomas A. White, Krzysztof Wrona, Chen Xu, Oleksandr Yefanov, Nadia Zatsepin, Jiaguo Zhang, Markus Perbandt, Adrian P. Mancuso, Christian Betzel, Henry Chapman, and Anton Barty. Megahertz serial crystallography. *Nature Communications*, 9(1):4025, 2018.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

[47] Matthew D. Hoffman, David M. Blei, Chong Wang, and John Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(4):1303–1347.

[48] Dan Foreman-Mackey, David W. Hogg, David S. Fulford, daft bot, László Dobos, Brian McFee, Kevin P. Murphy, Oliver Lindemann, Pierre Gerold, and Varun Agrawal. daft-dev/daft: daft v0.1.2.

[49] CCP4 and Global Phasing Ltd. Gemmi - library for structural biology [software]. `https://github.com/project-gemmi/gemmi`, 2020.

[50] J. D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9(3):90–95, 2007.

[51] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, September 2020.

[52] The pandas development team. pandas-dev/pandas: Pandas, February 2020.

[53] Jack B. Greisman, Kevin M. Dalton, and Doeke R. Hekstra. *reciprocalspaceship*: a Python library for crystallographic data analysis. *Journal of Applied Crystallography*, 54(5):1521–1529, Oct 2021.

[54] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[55] Michael L. Waskom. seaborn: statistical data visualization. *Journal of Open Source Software*, 6(60):3021, 2021.

[56] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[57] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. TensorFlow Distributions. *arXiv:1711.10604 [cs, stat]*, November 2017.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

APPENDIX

Kevin M. Dalton
Department of Molecular & Cellular Biology
Harvard University
Cambridge, MA 02138
kmdalton@fas.harvard.edu

Jack B. Greisman
Department of Molecular & Cellular Biology
Harvard University
Cambridge, MA 02138
greisman@g.harvard.edu

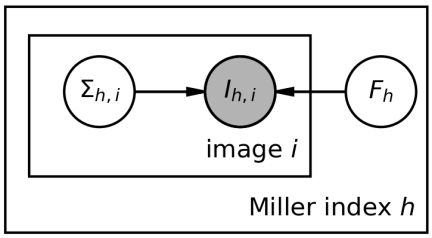
Doeke R. Hekstra
Department of Molecular & Cellular Biology
John A. Paulson School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138
doeke_hekstra@harvard.edu

September 7, 2022

1 Merging X-ray Data by Variational Inference

Observed reflection intensities can be thought of as the product of diffraction in an ideal experiment and a local scale which describes the systematic error in each reflection observation. This implies the graphical model,

Merging as a Graphical Model



Observed Data

- $I_{h,i}$: Intensities
- $\sigma_{I_{h,i}}$: Uncertainties

Latent Variables

- Σ : Reflection scales
- F_h : Structure factor amplitudes

Figure S1

with a corresponding joint distribution which factorizes as

$$p(I, F, \Sigma) = p(I|F, \Sigma)p(F)p(\Sigma).$$

In this setting, it is most desirable to estimate the posterior,

$$\begin{aligned} p(F, \Sigma|I) &= \frac{p(I, F, \Sigma)}{p(I)} \\ &= \frac{p(I|F, \Sigma)p(F)p(\Sigma)}{p(I)}. \end{aligned}$$

The exact posterior is generally intractable for such problems. So, we posit an approximate posterior q taken from a parametric family of distributions. This is the so-called variational distribution or surrogate posterior. We then use

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

optimization to learn parameters of q such that it approximates the desired posterior. One way to accomplish this is to minimize the Kullback-Leibler divergence between q and the posterior.

$$\begin{aligned}\mathcal{D}_{KL}\left[q||p(F,\Sigma|I)\right] &= \mathbb{E}_q\left[\log q - \log p(F,\Sigma|I)\right] \\ &= \mathbb{E}_q\left[\log q - \log p(I|F,\Sigma) - \log p(F) - \log p(\Sigma) + \log p(I)\right]\end{aligned}$$

Note that the expectation, $\mathbb{E}_q\left[\log p(I)\right]$ does not depend on the parameters of q. It is therefore a constant. Disregarding this constant term,

$$\mathcal{D}_{KL}\left[q||p(F,\Sigma|I)\right] \propto \mathbb{E}_q\left[\log q - \log p(I|F,\Sigma) - \log p(F) - \log(\Sigma)\right],$$

and negating,

$$ELBO(q) = \mathbb{E}_q\left[-\log q + \log p(I|F,\Sigma) + \log p(F) + \log(\Sigma)\right],$$

leads to the optimization objective of variational inference, which is called the Evidence Lower BOund (ELBO), [1, 2]. Maximizing this quantity with respect to the variational distribution, q, recovers an approximation to the posterior distribution. After re-arranging the terms,

$$\begin{aligned}ELBO\left(q\right) &= \mathbb{E}_q\left[\log p(I|F,\Sigma) - \log q + \log p(F) + \log p(\Sigma)\right] \\ &= \mathbb{E}_q\left[\log p(I|F,\Sigma)\right] - \mathbb{E}_q\left[\log q - \log p(F) - \log p(\Sigma)\right] \\ &= \mathbb{E}_q\left[\log p(I|F,\Sigma)\right] - \mathcal{D}_{KL}\left(q||p(F)p(\Sigma)\right),\end{aligned}$$

it is clear that the ELBO can be thought of as the sum of expected log-likelihood of the data and the Kullback-Leibler divergence between the surrogate posterior and the the prior. The KL divergence term acts as a penalty which discourages the surrogate posterior from wandering too far from the prior distribution. From the frequentist perspective, this is similar to a regularized maximum-likelihood estimator. This general form of the ELBO applies equally to any parameterization of the graphical model in Figure S1. The parameterization used in this work slightly simplifies this objective, as we'll show in the following section.

2 The parameterization used in Careless

The graphical model in Figure S1, implies that the prior distribution $p(F,\Sigma)$ factorizes $p(F)p(\Sigma)$. Therefore, it is convenient to assume the surrogate posterior,

$$q(F,\Sigma) = q_F(F)q_\Sigma(\Sigma),$$

consists of statistically independent distributions for F and Σ. This is a modeling choice, and it need not be the case in other variational merging models with this graph. Factorizing q leads to an $ELBO$,

$$ELBO\left(q\right) = \mathbb{E}_q\left[\log p(I|F,\Sigma)\right] - \mathcal{D}_{KL}\left(q_F||p(F)\right) - \mathcal{D}_{KL}\left(q_\Sigma||p(\Sigma)\right),$$

with separate Kullback-Leibler divergences for F and Σ. We can now begin to consider priors for each of these surrogate posteriors.

2.1 An Uninformative Prior on Scales

It is difficult to reason about the appropriate prior distribution for scales, $p(\Sigma)$. In all likelihood, this prior depends intimately on the details of the experiment. It will vary by sample and apparatus. In this work we choose an

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Model Parameters

Scale function

θ : MLP kernels and biases

Structure factors

$\mu_{q_{F_h}}$: location

$\sigma_{q_{F_h}}$: scale

Model Parameterization

$$F_h \sim q_h$$
$$\Sigma_{h,i} \sim q_{\Sigma_{h,i}}$$
$$I_{h,i} \sim \mathcal{N}ormal\left(F_h^2\Sigma_{h,i}, \sigma_{I_{h,i}}\right)$$
or
$$I_{h,i} \sim \mathcal{S}tudent\mathcal{T}\left(d.f., F_h^2\Sigma_{h,i}, \sigma_{I_{h,i}}\right)$$

Variational Distributions

$$q_{F_h} = \mathcal{T}runc\mathcal{N}ormal(\mu_{q_{F_h}}, \sigma_{q_{F_h}}, 0, \infty)$$
$$\mu_{\Sigma_{h,i}}, \sigma_{\Sigma_{h,i}} = f_\theta\left(M_{h,i}\right)$$
$$q_{\Sigma_{h,i}} = \mathcal{N}ormal\left(\mu_{\Sigma_{h,i}}, \sigma_{\Sigma_{h,i}}\right)$$

Priors

$$F_h \sim Wilson(h, \epsilon_h)$$
$$\Sigma_{h,i} \sim q_{\Sigma_{h,i}}$$

Figure S2: Summary of the model parameterization

uninformative prior, $\Sigma \sim q(\Sigma)$. Thereby, the second divergence term in the $ELBO$ becomes zero, and whatever parameters define q_Σ are simply allowed to optimize as dictated by the likelihood term. The objective used in this work is therefore

$$ELBO\left(q\right) = \mathbb{E}_q\left[\log p(I|F,\Sigma)\right] - \mathcal{D}_{KL}\left(q_F||p(F)\right).$$

2.2 Posterior Structure Factors

In this work, the surrogate posteriors of structure factors are independently parameterized by truncated normal distributions.

$$q_{F_h}(F) = \mathcal{T}runcated\mathcal{N}ormal\left(F|\mu_{F_h}, \sigma_{F_h}\right)$$

with location and scale parameters μ_{F_h} and σ_{F_h} and support

$$F \in \begin{cases} [0,\infty) & centric \\ (0,\infty) & acentric. \end{cases}$$

Both the location and scale parameters are constrained to be positive. In our work this constraint is implemented with the softplus function, $softplus(x) = \log(\exp(x) + 1)$.

2.3 Posterior Scales

As noted in section 2.1, we choose not to impose a structured prior on reflection scales. Rather, we assert that the scale of a reflection should be computable from the geometric metadata recorded about each reflection during integration. Therefore, our model infers a function which ingests metadata and outputs scale factors (Figure S3).

By default, we parameterize this function as a deep neural network with parameters θ. In particularly, we use a multilayer perceptron with leaky rectified linear units (ReLU) as the activation. The parameters, correspond to the kernels and biases of each layer. The kernels are initialized to the identity matrix and biases to zeros. The number of hidden units in each layer takes the width of the metadata by default, but this is user-configurable. We set the default depth of the neural network to twenty layers which we find offers a reasonable balance of performance and stability. The model has a final, linear layer with 2 units. The output of the last layer is interpreted as the mean and standard deviation of the scale distribution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

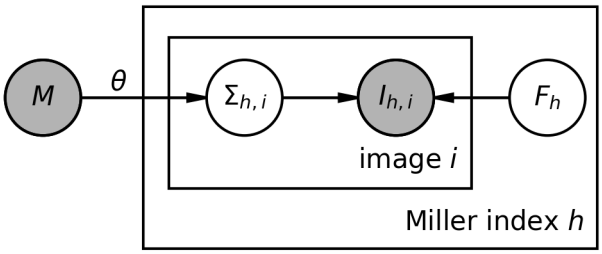


Figure S3: Graphical model with a scale function

$$\begin{aligned}\left[\mu_{\Sigma_{h,i}}, \sigma_{\Sigma_{h,i}}\right] &= f_\theta(M_{h,i}) \\ q_{\Sigma_{h,i}} &= \mathcal{N}ormal\left(\mu_{\Sigma_{h,i}}, \sigma_{\Sigma_{h,i}}\right)\end{aligned}$$

The correct scale function is the one which allows the model to recapitulate the data while letting the structure factors follow the desired prior distribution. Provided rich enough metadata about each reflection observation, variational inference will recover such a function.

One must use caution when selecting metadata. If information about the reflection intensities is provided to the scale function, the scale function may bypass the structure factors to directly minimize the expected log likelihood. This leads to poor structure factor estimates. We recommend against including data such as the reflection uncertainties in the metadata, as they are strongly correlated with the intensities.

2.4 Wilson's Priors

Wilson's priors,

$$p(F_h) = \mathcal{W}ilson(F_h) = \begin{cases} \mathcal{H}alfnormal\left(F_h|\epsilon_h\right) = \sqrt{\frac{2}{\pi\epsilon_h}}\exp\left(-\frac{F_h^2}{2\epsilon_h}\right) & h \text{ is centric} \\ \mathcal{R}ayleigh\left(F_h|\epsilon_h\right) = \frac{2}{\epsilon}F_h\exp\left(-\frac{F_h^2}{\epsilon_h}\right) & h \text{ is acentric,} \end{cases} \tag{1}$$

express the expected distributions of structure factors under the assumption that atoms are uniformly distributed within the unit cell[3]. The probability distribution over the structure factor amplitude F_h with Miller index h is expressed in terms of the multiplicity of the reflection ϵ_h. The multiplicity, a feature of the crystal's space group, is a constant which can be determined for each Miller index. It corresponds to the contribution to the relative intensity of each reflection solely due to crystal symmetry. The Wilson prior has separate parameterizations for centric and acentric reflections. This form of Wilson's priors differs from the one employed in the French Wilson algorithm [4] in that it is independent of the scale, Σ. Because of this choice, the scale function can be inferred in parallel with the structure factor amplitudes. However, it implies that the structure factors output by Careless are on the same scale across resolution bins. This is an important consideration for downstream processing. Careless output may, for some applications, need to be rescaled to meet the expectations of crystallographic data analysis packages.

2.5 Likelihood functions

The first term in the $ELBO$ is an expected log likelihood. In this work, we present two parameterizations of this term: a normal distribution

$$p(I_{h,i}|F_h, \Sigma_{h,i}) = \mathcal{N}ormal(I_{h,i}|F_h^2\Sigma_{h,i}, \sigma_{I_{h,i}}),$$

which is suitable for data with few outliers, as well as a robust t-distribution

$$p(I_{h,i}|F_h, \Sigma_{h,i}) = \mathcal{S}tudent\mathcal{T}(I_{h,i}|d.f., F_h^2\Sigma_{h,i}, \sigma_{I_{h,i}}),$$

which adds an additional hyperparameter. The degrees of freedom, $d.f.$, titrates the robustness of the model toward outliers. In the limit $d.f. \to \infty$, the t-distribution approaches a normal distribution.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

2.6 Model training

In this work, we use the reparameterization trick to estimate gradients of the $ELBO$ with respect to the parameters of the variational distributions, q. First applied in the Variational Autoencoder [5], reparameterization is a common tool to estimate gradients of probabilistic programs. In our implementation, the $ELBO$ is approximated by random samples from the surrogate distributions,

$$\begin{aligned} ELBO\left(\mu_{q_F}, \sigma_{q_F}, \theta\right) &= \mathbb{E}_q\left[\log p\left(I|F^2\Sigma\right)\right] - \mathcal{D}_{KL}(q_F||p(F)) \\ &\approx \sum_{s=1}^{S}\left[\sum_i \sum_h \log p(I_{h,i}|F_{h,s}^2\Sigma_{h,i,s}, \sigma_{I_{h,i}}) - \sum_h \left(\log q_{F_h}(F_{h,s}) - \log p(F_{h,s})\right)\right] \end{aligned}$$

where $F_{h,s}$ and $\Sigma_{h,i,s}$ denote reparameterized samples from the surrogate posteriors, and the number of Monte Carlo samples, S is a hyperparameter. By default a single sample is used ($S = 1$). For training, we use the Adam optimizer [6] with hyperparameters $\alpha = 0.001$, $\beta_1 = 0.9$, and $\beta_2 = 0.99$.

2.7 Cross-validation

Careless provides two modes of cross-validation. In the first paradigm, the model is first trained on the full data set yielding structure factor estimates and neural network weights. Next, the data are partitioned randomly into halves by image. Using the neural network weights learned from the full data set, each half is merged separately by optimizing the structure factors. During this process the neural network weights remain fixed. The resulting pair of structure factor estimates may be correlated to produce a measure similar to the canonical $CC_{1/2}$ widely used in crystallography. This mode of cross-validation does not necessarily inform the user about the degree of overfitting. Rather, the $CC_{1/2}$ value is more indicative of the data consistency.

The second type of cross-validation supported by Careless is intended to explicitly test for overfitting in the scale function. In this mode, a fraction of the data is held out during training time. After training, the model is applied to these data in order to predict intensities for the held out fraction. The correlation between the observed intensities and the predictions provides an estimate for how well the model generalizes. The choice of summary statistic is up to the user. However, we recommend Spearman's rank correlation coefficient as a robust alternative to Pearson's. In the following section, we address the issue of how to recover intensity predictions and moments from our model.

2.8 Predictions

Model predictions are essential to quantify model overfitting by cross-validation. The predicted reflections implied by our model are the product of random variables,

$$\hat{I}_{h,i} = \Sigma_{h,i} F_h^2,$$

which is itself a distribution for which we do not have an analytical expression. It is however possible to compute the expected value,

$$\begin{aligned} \langle \hat{I}_{h,i} \rangle &= \langle \Sigma_{h,i} \rangle \langle F_h^2 \rangle \\ &= \langle \Sigma_{h,i} \rangle \left(\langle F_h \rangle^2 + \sigma_{F_h}^2 \right). \end{aligned}$$

The first term in the product,

$$\begin{aligned} \langle \Sigma_{h,i} \rangle &= \mu_{\Sigma_{h,i}} \\ &= f_\theta(M_{h,i})_1, \end{aligned}$$

is computed by the scale function, f_θ, from the metadata vector, $M_{h,i}$. The second term, is calculated from the moments of the truncated normal surrogate posterior, q_{F_h}. These moments have analytical expressions which are implemented in many statistical libraries including TensorFlow Probability[7] which we use in this work.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

It is also possible to compute the second moment of the predictions,

$$\begin{aligned}\sigma^2_{\hat{I}_{h,i}} &= \langle \hat{I}^2_{h,i} \rangle - \langle \hat{I}_{h,i} \rangle^2 \\ &= \langle F^4_h \rangle \langle \Sigma^2_{h,i} \rangle - \langle \hat{I}_{h,i} \rangle^2 \\ &= \langle F^4_h \rangle \left(\sigma^2_{\Sigma_{h,i}} + \mu^2_{\Sigma_{h,i}} \right) - \langle \hat{I}_{h,i} \rangle^2,\end{aligned}$$

where the fourth non-central moment, $\langle F^4_h \rangle$ of q_{F_h} has an analytical expression which is implemented in SciPy [8].

2.9 Harmonic deconvolution for Laue diffraction

To implement harmonic deconvolution, the Careless ELBO approximator needs to be modified to update the center of the likelihood distribution. By summing over each contributor on the central ray, the new ELBO approximation becomes

$$ELBO\left(\mu_{q_F}, \sigma_{q_F}, \theta\right) \approx \sum_{s=1}^{S} \left[\sum_i \sum_{ray} \log p \left(I_{ray,i} \middle| \sum_{h \in ray} \left(F^2_{h,s} \Sigma_{h,i,s} \right), \sigma_{I_{ray,i}} \right) - \sum_h \left(\log q_{F_h}(F_{h,s}) - \log p(F_{h,s}) \right) \right]$$

which is readily optimized by same protocol demonstrated in Figure S5. In the code base, harmonics are handled by having a separate class of likelihood objects for Laue experiments. In practice, one could use the polychromatic likelihood to merge monochromatic data with no ill effect on the quality of the results. In that sense, this is the more general version of the ELBO for diffraction data. However, doing so would incur a performance cost given the underlying implementation, which is why we maintain separate likelihoods for mono and polychromatic experiments. Regardless, the core merging class inside Careless is competent to fit both sorts of data.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

3 Supplementary Information

3.1 Supplementary Note: Summary of conventional X-ray diffraction data processing

For reference, we summarize the conventional workflow for conversion of diffraction images to a non-redundant set of structure factor amplitudes—a process known as data reduction. This involves sequential application of a series of algorithms. First, during indexing, the orientation of the crystal lattice is determined, starting from a subset of strong reflections in each image, mapping each observed reflection to its Miller index ((h, k, l), abbreviated here as h). Next, in geometry refinement the mismatch between predicted and observed reflection centroids is minimized by optimizing estimates of detector position, beam center, crystal unit cell parameters, goniometer rotation axis, and other geometric parameters. In integration, the intensities of the reflections are then estimated, either by summing the pixel values in a region around the predicted spot centroid or by fitting a profile model to each reflection. In either case the region around a reflection can be used to estimate the background contribution from X rays resulting from other processes, such as inelastic scatter and scatter by sample mounts, bulk liquid, or air. This background is typically subtracted from the raw integrated intensity to yield an intensity estimate for each predicted reflection.

Next, the integrated intensities need to be converted into structure factor amplitudes. Traditionally this is done in three steps. In scaling, scale parameters are learned per image to account for beam intensity fluctuations, crystal disorder, radiation damage, sample absorption, and other effects which vary systematically throughout a diffraction experiment. The parameters are applied to each reflection to yield scaled intensities. At this point, the data still contain many redundant observations for each Miller index. In merging, equivalent observations are merged by weighted averaging. This step assumes that the errors in the reflection intensities are normally distributed. This works well if the errors in the reflection intensities are dominated by photon counting statistics or other random additive errors, but is sensitive to outliers. In French-Wilson scaling, the merged intensities are then "corrected" because they can be negative, resulting from errors in the estimated intensity of reflections and background, and inconsistent with the reflection intensities being proportional to squared structure factor amplitudes, and therefore positive.

In this last step, the Wilson distribution [3] serves as a prior probability distribution, or prior. This prior is combined with a statistical model of the true intensity given the observed merged intensity to yield a posterior probability distribution, or posterior, of the true merged intensity. This true intensity, the squared structure factor amplitude, is guaranteed to be positive. The Wilson distribution is parametrized by the mean intensity which is usually modeled as resolution-dependent.

3.2 Data collection and analysis

Data collection for hen egg white lysozyme was described in ref. [9]. Data for photoactive yellow protein were collected as described in ref. [10]. Collection of thermolysin data was described in [11]. Scaling and merging was performed using Careless version 0.2.0. DIALS version 3.1.4 was used to index and integrate observed reflections for hen egg white lysozyme. Aimless version 0.7.4 was used to merge the integrated intensities for hen egg white lysozyme data. Precognition version 5.2.2 was used to index and integrate the polychromatic PYP data. cctbx.xfel version 2021.11.dev3+4.g05389c3054 was used to scale and merge thermolysin XFEL data; the merging parameters are available in the Zenodo deposition. All model refinement and phasing was performed in PHENIX version 1.18.2. The refinement outputs and log files including parameter settings are deposited in Zenodo. The anomalous peak heights presented in Figure 5 and Table S2, were quantified using the "Difference map peaks..." function Coot version 0.9.6.

4 Supplementary Figures and Tables

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

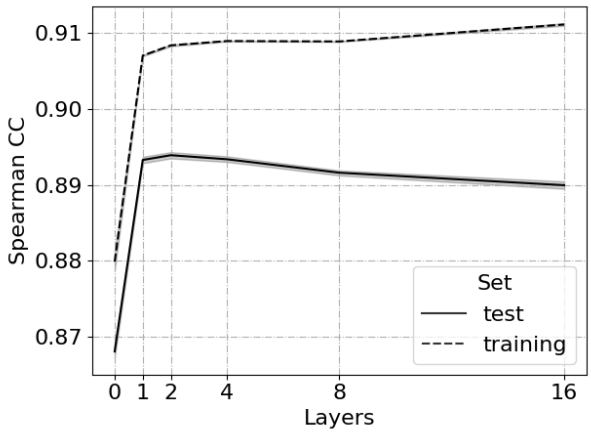


Figure S4: Selection of the number of image layers by 10-fold cross-validation of observed against predicted intensities for merging thermolysin serial crystallography.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.


```
steps=10000
n_layers = 20
mc_samples = 3
p_centric  = tfd.HalfNormal(np.sqrt(multiplicity))
p_acentric = tfd.Weibull(2., np.sqrt(multiplicity))

#Construct variational distributions
loc_init   = tf.where(centric, p_centric.mean(), p_acentric.mean())
scale_init = tf.where(centric, p_centric.stddev(), p_acentric.stddev())
q = tfd.TruncatedNormal(
    loc = tf.Variable(loc_init),
    scale = tfp.util.TransformedVariable(scale_init, tfp.bijectors.Softplus()),
    low = tf.where(centric, 0., 1e-30),
    high = 1e30,
)

#Construct error model
likelihood = tfd.Normal(loc=intensities, scale=uncertainties)

#Construct scale function
n,d = metadata.shape
NN = tf.keras.models.Sequential()
NN.add(tf.keras.Input(d))
for i in range(n_layers):
    NN.add(tf.keras.layers.Dense(d, kernel_initializer='identity'))
NN.add(tf.keras.layers.Dense(2, kernel_initializer='identity'))

#Evaluate the elbo
def minus_elbo():
    z = q.sample(mc_samples)
    F = tf.gather(z, miller_id, axis=1)
    loc, scale = tf.unstack(NN(metadata), axis=1)
    Sigma = tfd.Normal(loc, scale).sample(mc_samples)
    log_likelihood = tf.reduce_sum(likelihood.log_prob(F * F * Sigma))
    log_p_z = tf.where(centric, p_centric.log_prob(z), p_acentric.log_prob(z))
    log_q_z = q.log_prob(z)
    kl_div = tf.reduce_sum(log_q_z - log_p_z)
    return -log_likelihood + kl_div

#Train the model
optimizer = tf.keras.optimizers.Adam()
for i in range(steps):
    optimizer.minimize(minus_elbo, [q.trainable_variables , NN.trainable_variables])

#Export the results
F,SigF = q.mean().numpy(), q.stddev().numpy()
```

Figure S5: **Example implementation of the Careless model using TensorFlow Probability**. This example code is fully functional and available as a script on the Careless GitHub page [12].

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

Degrees of Freedom	Figure of Merit	Bayes-CC $\pm$ S.D.	Sites
aimless	0.45	27.3 $\pm$ 14.4	11
0.125	0.23	9.9 $\pm$ 12.3	10
0.25	0.19	15.8 $\pm$ 14.0	10
0.5	0.4	25.7 $\pm$ 14.4	11
1	0.21	13.3 $\pm$ 13.3	10
2	0.42	32.1 $\pm$ 13.7	10
4	0.43	31.6 $\pm$ 13.6	11
8	0.43	31.8 $\pm$ 13.6	10
16	0.42	33.8 $\pm$ 13.4	11
32	0.43	26.6 $\pm$ 14.6	11
64	0.43	32.6 $\pm$ 13.3	10
128	0.41	33.4 $\pm$ 13.3	10
256	0.42	28.3 $\pm$ 14.1	10
512	0.41	30.1 $\pm$ 13.9	10
1024	0.39	27.0 $\pm$ 14.2	11
∞	0.39	24.3 $\pm$ 14.5	11

Table S1: *Ab initio* phasing results from AutoSOL using thorough defaults and searching for 10 sulfur sites. The Figure of Merit of phasing places an upper bound on the quality of an experimentally phased electron density map. The Bayes-CC is an effective estimate of the quality of an experimental map before density modification based on its skew. See [13] for further description of both metrics.

Site	Aimless	Careless (16 d.f.)	Careless (∞ d.f.)	Careless (image layers)	Careless (transfer)
Cys-6	14.91	11.97	10.45	14.48	15.86
Met-12	17.52	10.92	9.8	16.48	17.33
Cys-30	19.75	14.63	12.10	19.47	20.51
Cys-64	17.59	14.01	9.75	18.02	19.06
Cys-76	14.84	11.39	10.85	13.71	15.82
Cys-80	18.29	15.30	12.88	18.63	20.90
Cys-94	14.98	12.28	8.86	14.54	15.95
Met-105	19.89	14.84	11.67	20.30	21.48
Cys-115	19.38	16.13	11.41	18.92	20.29
Cys-127	14.62	10.64	10.16	14.50	15.35

Table S2: Anomalous omit map peak heights from PHENIX refinement with isotropic b-factors and rigid body refinement. The careless merging options used here are described in detail on the careless-examples GitHub page.

bioRxiv preprint doi: https://doi.org/10.1101/2021.01.05.425510; this version posted September 8, 2022. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC 4.0 International license.

References

[1] Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine Learning*, 37(2):183–233, November 1999.

[2] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 112(518):859–877, April 2017. arXiv: 1601.00670.

[3] A. J. C. Wilson. The probability distribution of x-ray intensities. *Acta Crystallographica*, 2(5):318–321, 1949.

[4] S. French and K. Wilson. On the treatment of negative intensity observations. *Acta Crystallographica Section A*, 34(4):517–525, 1978.

[5] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, May 2014.

[6] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization. *arXiv:1412.6980 [cs]*, January 2017.

[7] Joshua V. Dillon, Ian Langmore, Dustin Tran, Eugene Brevdo, Srinivas Vasudevan, Dave Moore, Brian Patton, Alex Alemi, Matt Hoffman, and Rif A. Saurous. TensorFlow Distributions. *arXiv:1711.10604 [cs, stat]*, November 2017.

[8] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[9] Jack B. Greisman, Kevin M. Dalton, Candice J. Sheehan, Margaret A. Klureza, and Doeke R. Hekstra. Native SAD phasing at room temperature. *bioRxiv*, page 2021.12.13.472485.

[10] M. Schmidt, V. Srajer, R.H. Henning, H. Ihee, N. Purwar, J. Tenboer, and S. Tripathi. Protein energy landscapes determined by five-dimensional crystallography. *Acta Crystallographica Section D Biological Crystallography*, 69:2534–2542, 2013.

[11] Jan Kern, Rosalie Tran, Roberto Alonso-Mori, Sergey Koroidov, Nathaniel Echols, Johan Hattne, Mohamed Ibrahim, Sheraz Gul, Hartawan Laksmono, Raymond G. Sierra, Richard J. Gildea, Guangye Han, Julia Hellmich, Benedikt Lassalle-Kaiser, Ruchira Chatterjee, Aaron S. Brewster, Claudiu A. Stan, Carina Glöckner, Alyssa Lampe, Dörte DiFiore, Despina Milathianaki, Alan R. Fry, M. Marvin Seibert, Jason E. Koglin, Erik Gallo, Jens Uhlig, Dimosthenis Sokaras, Tsu-Chien Weng, Petrus H. Zwart, David E. Skinner, Michael J. Bogan, Marc Messerschmidt, Pieter Glatzel, Garth J. Williams, Sébastien Boutet, Paul D. Adams, Athina Zouni, Johannes Messinger, Nicholas K. Sauter, Uwe Bergmann, Junko Yano, and Vittal K. Yachandra. Taking snapshots of photosynthetic water oxidation using femtosecond X-ray diffraction and spectroscopy. *Nature Communications*, 5(1):4371, July 2014.

[12] Kevin M. Dalton and Jack B. Greisman. Hekstra-Lab/careless, November 2020. https://github.com/Hekstra-Lab/careless.

[13] T. C. Terwilliger, P. D. Adams, R. J. Read, A. J. McCoy, N. W. Moriarty, R. W. Grosse-Kunstleve, P. V. Afonine, P. H. Zwart, and L.-W. Hung. Decision-making in structure solution using Bayesian estimates of map quality: the PHENIX AutoSol wizard. *Acta Crystallographica Section D: Biological Crystallography*, 65(6):582–601, June 2009.