CARELESS: A VARIATIONAL BAYESIAN MODEL FOR MERGING X-RAY DIFFRACTION DATA

A PREPRINT

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

January 6, 2021

ABSTRACT

The primary data in X-ray diffraction experiments are images which contain bright spots known as reflections. In order to compute the electron density of the crystal which yields these reflections, the intensity of each reflection must be estimated. Redundant observations must be parsimoniously reduced to a set of summary intensities and corresponding error estimates. This task is confounded by a range of systematic effects leading to equivalent reflections being expressed on different scales. Here, we present a modern Bayesian solution to this problem, `careless`, which uses deep learning and variational inference to simultaneously merge and scale reflection observations. We successfully apply this method to conventional monochromatic X-ray diffraction data, polychromatic diffraction data, and finally X-ray free-electron laser data. Thereby, we find that this single model is generally applicable to many types of X-ray diffraction experiments.

***K*eywords** X-ray crystallography · Variational inference · Deep learning

1 Introduction

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, or reflections, with intensities proportional to the squares of the amplitudes of the corresponding Fourier components, or structure factors, of the electron density in the crystal. Each structure factor reports on a specific 3D direction and spatial frequency, indexed by triplets of discrete 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, X-ray diffraction has revolutionized our understanding of the molecular basis of life over the past 60 years.

Recent advances in X-ray diffraction now enable direct visualization of biomolecules in action, imaged by short X-ray pulses generated at synchrotrons and X-ray Free-Electron Lasers (XFELs). Full realization of the promise of these methods requires very accurate measurement of the often small changes in structure factor amplitudes that result from conformational change. In addition, the conditions of these new time-resolved experiments are complex and rapidly evolving—including the use of polychromatic X-ray beams, new detectors, new methods for sample delivery, and a range of perturbation methods. Spurred on by these challenges, we here introduce a new approach to accurately scale

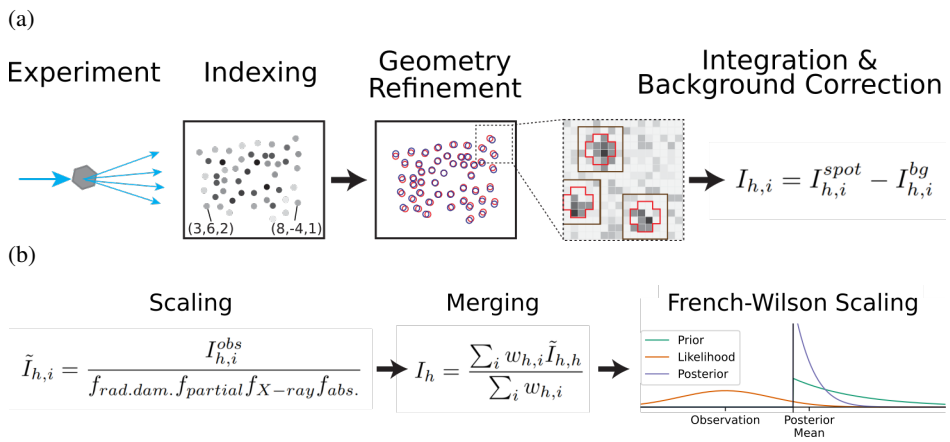


Figure 1: Classical crystallographic data reduction. (a) X-rays diffracted from a crystal are collected. In indexing, the Miller index of each spot is assigned after finding approximate unit cell dimensions and crystal orientation. Geometric parameters are refined, minimizing the mismatch between predicted (blue circles) and observed (red) spot centroids. The total X-ray intensity is integrated over predicted spot pixels (red box), and background is estimated based on surrounding pixels. (b)Scaling corrects effects of various physical factors, followed by weighted averaging of corrected intensities. French-Wilson scaling then yields positive estimates.

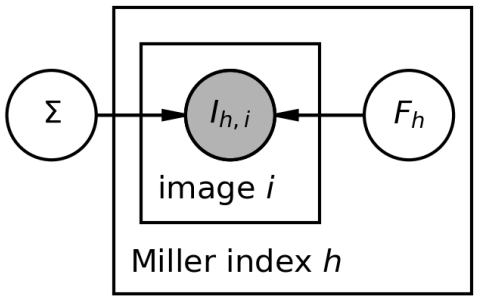
and merge X-ray reflections in order to extract such subtle structural signals. We refer to this approach as `careless`, in homage to earlier methods[1]. `careless` is based on variational inference, making it highly flexible in terms of its physical and statistical model, and providing a unifying statistical treatment of XFEL and synchrotron data, both poly- and monochromatic.

Conversion of diffraction images to a non-redundant set of structure factor amplitudes—a process known as data reduction—involves sequential application of a series of algorithms (Fig. 1a). 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 a Miller index. Next, in geometry refinement the mismatch between predicted and observed reflection centroids is minimized by optimizing estimates of detector position, tilt, 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 the intensity estimate associated with each predicted reflection.

At this point, the integrated intensities need to be converted into structure factor amplitudes (Fig. 1b). 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 any other effects which may 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 a weighted summation. The weighted summation relies on the assumption that the errors affiliated with each reflection observation are normally distributed. This works well if the errors in the observations 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 random fluctuations and errors in the estimated intensity of reflections and background, inconsistent with the reflection intensities being proportional to squared structure factor amplitudes, and therefore positive.

This last step follows a Bayesian argument: structure factor amplitudes are positive and can, under simple assumptions, be expected to follow the so-called Wilson distribution [2]. In a Bayesian sense, the Wilson distribution can serve as a precise prior probability distribution, or prior, which can be 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, or squared structure factor amplitude, which is guaranteed to be positive. The Wilson distributions are parameterized by the mean intensity, Σ. In an ideal crystal this mean intensity would be a simple constant. In real crystals, Σ cannot be taken as constant across all reflections. In particular crystal disorder often manifests as a resolution-dependent decrease in mean intensity. The Bayesian routine introduced by French & Wilson first estimates Σ per resolution shell in

(a) Probabilistic Graphical Model



Σ Reflection scale factor

F_h Structure factor amplitude

$I_{h,i}$ Observed reflection intensity

(b) Intensity Generating Model

$$\Sigma \sim P(\Sigma)$$
$$F_h \sim P(F_h)$$
$$I_h^i \sim \mathcal{L}\left(F_h^2 \Sigma_{h,i}, \sigma_{I_{h,i}}\right)$$

(c) Joint Model

$$P(I, F, \Sigma) = P(I|F, \Sigma)P(\Sigma)P(F)$$
$$P(F) = \prod_h P(F_h)$$
$$P(I|F, \Sigma) = \prod_{h,i} \mathcal{L}(I|F_h^2 \Sigma_{h,i}, \sigma_{I_{h,i}})$$

Figure 2: (a) General graphical model for merging X-ray data showing the statistical dependencies between the random variables underlying the reflections observed in the data. Observed intensities, $I_{h,i}$ for image i and Miller index h are dependent on the structure factor, F_h, and the scale factor, Σ. (b) The general parameterization underlying the graph in panel (a). Σ and F_h are drawn from unspecified prior distributions. $I_{h,i}$ is drawn from a family likelihood distribution centered at $F_h^2 \Sigma_{h,i}$ with empirical standard deviation $\sigma_{I_{h,i}}$. (c) The joint probability for a data set can be factorized over all variables in the graph.

reciprocal space, followed by interpolation to yield estimates for each reflection. The appropriate Σ for each reflection observation is then used to parameterize the prior distribution and calculate a posterior estimate of the merged intensity which is now strictly positive.

`careless` builds on this powerful idea, performing scaling, merging, and French-Wilson scaling in a single step—that is, it directly infers structure factor amplitudes from unscaled intensities. The probability calculation is "forward", predicting integrated intensities from structure factor amplitudes and experimental metadata. In this approach, optimization takes the place of estimation. As a consequence, analytical tractability of the inference is no longer a concern and the model relating structure factors to integrated intensities can be arbitrarily complex.

2 Merging X-ray Data by Variational Inference

2.1 Merging as a Graphical Model

The merging of X-ray data can be abstractly expressed by a probabilistic graphical model. The most general form of this model is illustrated in Figure 2a. This graph represents a statistical model which predicts the observed intensities $I_{h,i}$, for Miller index h in image i in the data set from structure factor amplitudes F_h and scale factors Σ, which must both be estimated. The structure factor amplitudes can be reasonably considered statistically independent across Miller indices. However, most contributions to scale factors vary slowly across the data set and are best accounted for using a global parameterization.

As uncertainty estimates, $\sigma_{I_{h,i}}$, are usually available for each intensity observation, the elements of the graph can be parameterized as in Figure 2, with $\mathcal{L}$ corresponding to a suitable error distribution. `careless` implements several choices of error distribution and structure factor prior distributions $P(F_h)$.

The goal of merging is to estimate the posterior distributions over structure factors, $P(F_h|I_{h,i}, \sigma_{I_{h,i}})$. Exact inference of this model presents two major obstacles. First, the posterior over structure factor amplitudes is high dimensional. Typically tens to hundreds of thousands of structure factor amplitudes need to be estimated to produce a high-resolution structure of a biomolecular crystal. Exact inference of such a high-dimensional posterior distribution is beyond the scope of routine computing. Instead, we will address this by variational inference, as described in the next section. Second, there is no obvious choice of prior distribution on scale factors, $P(\Sigma)$, which is a prerequisite for exact inference. Indeed, the appropriate form and parameters the scale factor prior distribution depends heavily on the particulars of the experiment which vary between facilities and samples. Fortunately, the success of crystallographic scaling models suggests that estimation of a scale function is well posed for each particular data set. Therefore, `careless` neglects

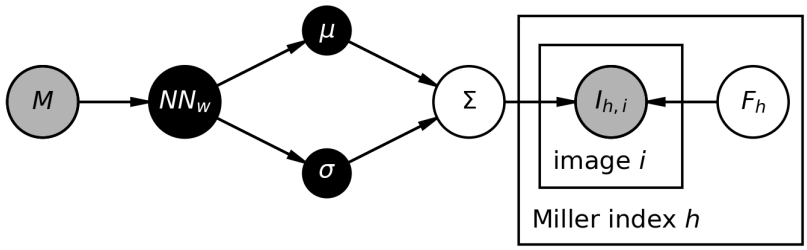
to impose a prior distribution, $P(\Sigma)$. Variational methods for approximate inference readily admit such parameters, allowing them to "float" by maximum likelihood in the absence of explicit prior distributions.

2.2 The `Careless` Parameterization

At heart, `careless` is based on a probabilistic forward model of X-ray diffraction: a stochastic description of how the data are generated. Such a forward calculation is not constrained by whether the inverse calculation is analytically tractable, and can include both explicit physics calculations and machine learning concepts. `Careless` learns two objects, the first of which is a bank of merged structure factor amplitudes and corresponding uncertainties. The second is a scale function that ingests metadata about each reflection observation and outputs the appropriate scale factor distribution for an observation. For example, these metadata may include information such as position on the detector, reflection resolution, goniometer rotation angle, or the observed Miller index of the reflection. The software can leverage any information about the observation which might be useful in determining the appropriate scale. The particular parameterization of the scale function used in `careless` is a deep, fully connected neural network, but other models can be implemented. `careless` then uses the scale function, reflection metadata, and structure factors to predict an intensity for each reflection observation in the data set. The scale function is learned such that the predictions match the data as best as possible while imposing a prior distribution on the structure factors. In `careless`, the intensity, $I_{h,i}$, corresponding to Miller index h and image i must have an empirical error estimate, $\sigma_{I_{h,i}}$. Each reflection observation must be associated with a metadata vector, $m_{h,i} \in \mathbb{R}^d$. The metadata are standardized by `careless` to stabilize the optimization. The graphical model implemented in `careless`, parameterizes the scale factors Σ in the general model (Figure 2) by a neural network, NN_w with weights denoted by w which takes as input the metadata vectors. With some abuse of notation, the model is schematized in a graphical model in Figure 3a.

During training, `careless` attempts to estimate posterior distribution over structure factor amplitudes corresponding to each non-redundant Miller index observed in the experiment. To do so, the model relies on *variational inference*[3, 4]. In this approach, the exact density of the joint posterior distribution is not determined, as in, for example, Markov Chain Monte Carlo sampling. Instead, the shape is assumed. By default, the assumed posterior for the scale factors are normal distributions with means and variances dictated by the neural network (Fig. 3b), while the posterior distribution for the structure factors is assumed to the product of truncated normal distributions on $[0, \infty)$ with independent location and scale parameters, μ_{q_h} and σ_{q_h} (Fig. 3c).

(a) Augmented graphical model



(b) Model

$$\begin{aligned} \mu_\Sigma, \sigma_\Sigma &= NN_w(M) \\ \Sigma &\sim \mathcal{N}(\mu_\Sigma, \sigma_\Sigma \mathbf{I}) \\ F_h &\sim P(F_h) \\ I_{h,i} &\sim \mathcal{L}\left(F_h^2 \Sigma_{h,i}, \sigma_{I_{h,i}}\right) \end{aligned}$$

(c) Variational approximation

$$\begin{aligned} q_h &= TruncatedNormal(\mu_{q_h}, \sigma_{q_h}) \\ \mu_\Sigma, \sigma_\Sigma &= NN_w(M) \\ \Sigma &\sim \mathcal{N}(\mu_\Sigma, \sigma_\Sigma \mathbf{I}) \\ F_h &\sim q_h(\mu_{q_h}, \sigma_{q_h}) \\ I_{h,i} &\sim \mathcal{L}\left(F_h^2 \Sigma_{h,i}, \sigma_{I_{h,i}}\right) \end{aligned}$$

Figure 3: The `careless` parameterization. (a) The `careless` augmented graphical model with the crystallographic scale factors, Σ, parameterized by a neural network transformation, NN_w with weight variables w, of the reflection metadata, M (b) The statiscal model generating reflection observations from metadata, M, and the structure factors, F_h. `Careless` does not place an explicit prior on the neural network parameters, w, or the scales, Σ. They are learned by maximum likelihood. (c) The variational approximation used by `careless`. The form of the posterior over structure factors is assumed truncated normal with support on $[0, \infty)$.

2.3 Implementation

Variational inference models are trained by maximizing the Evidence Lower BOund (ELBO) [4]. Training models by maximizing the ELBO is now routine and a number of straightforward and general methods [5, 6] for doing so have been proposed. `careless` relies on the method know as the reparameterization trick [6]. Reparameterization states the sampling operation as a transformed draw from a well-behaved distribution such as a standard normal. This allows gradients of expectations with respect to a random variable to be computed automatically under the assumption of a particular parameterization. Thereby, sampling operations can be included as a member of a computational graph and be efficiently optimized with modern machine learning libraries.

For `careless`, this takes the following form:

$$\begin{aligned} ELBO(\mu_{q_h}, \sigma_{q_h}, w) &= \sum_{h,i} \mathbb{E}\left[\log \mathcal{L}(I_{h,i}|F_h^2\Sigma_{h,i}, \sigma_{I_{h,i}})\right] - \sum_h D_{KL}\left(q_h(F_h)||P(F_h)\right) \\ &= \sum_{h,i} \mathbb{E}\left[\log \mathcal{L}(I_{h,i}|F_h^2\Sigma_{h,i}, \sigma_{I_{h,i}})\right] - \sum_h \mathbb{E}\left[\log q_h(F_h) - \log P(F_h)\right] \\ &\approx \sum_{j=1}^{s} \left[\sum_{h,i} \log \mathcal{L}(I_{h,i}|F_{h,j}^2\Sigma_{h,i,j}, \sigma_{I_{h,i}}) - \sum_h \left[\log q_h(F_{h,j}) - \log P(F_{h,j})\right]\right] \end{aligned} \tag{1}$$

where sampling is indicated by the index j. `Careless` relies on the reparameterized distributions implemented in Tensorflow Probability [7] to estimate the ELBO, and automatic differentiation as well as optimizers from Tensorflow [8]. Because these reparameterized gradients can be noisy, the loss function is optimized using the Adam optimizer [9] which was developed specifically for stochastic objective functions.

The entire probabilistic program can be implemented concisely in Python (Figure 4). The inputs to the algorithm are the per-reflection metadata, intensities, and uncertainty estimates. A proper implementation also requires a zero indexed vector of integer Miller index IDs, a vector indicating which indices are centric, and a vector of multiplicities. The example code in Figure 4 is fully functional, and is incorporated into an example script on the `careless` GitHub page [10].

2.4 Prior Distributions in `Careless`

`careless` implements several prior distributions for structure factor amplitudes, the most generally applicable being the Wilson distributions [2]. Wilson's priors,

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

arise from a model of random atomic coordinates inside the crystal's unit cell. The probability distribution over the structure factor amplitude F_h for 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. Briefly, it corresponds to the contribution to the relative intensity of each reflection solely due to crystal symmetry. The Wilson prior has separate parameterizations for the so-called centric and acentric reflections. This form of Wilson's priors differs from the one employed in the French Wilson algorithm [11] 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 Robust likelihoods, not outlier rejection

Traditional crystallography software relies on least squares routines for fitting scaling models. This paradigm asserts that the residuals during optimization must be well approximated by a normal distribution. Otherwise, the scale parameters will not converge to reasonable values. Most crystallographic data sets contain several outliers which would distort the inferred scales if least squares was applied naïvely. The traditional solution to this problem has been to intersperse rounds of scaling and statistical outlier rejection. This is a principled way to handle outliers in an optimization context

```python
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 4: Example implementation of the `Careless` model using TensorFlow Probability

which has been extremely successful for crystallography. In the context of variational inference, it is trivial to achieve a similar result by the likelihood function $\mathcal{L}$. The default likelihood function in `careless` is the probability density of the normal distribution corresponding to a normally distributed error model. However, `careless` also supports the use of Student's T distribution as the likelihood. This distribution is a popular choice for robust estimation as it has heavier tails than the normal distribution. The version of T distribution used in this work is a member of the location-scale family, and has 3 parameters: location, scale, and degrees of freedom. It includes several other distributions as limits or special cases. Notably, the T distribution with a single degree of freedom is equivalent to the Cauchy distribution. The limit of the T distribution as the degrees of freedom approach infinity is the normal distribution. The key to successfully applying the T-dsitributed error model, is to choose the degrees of freedom to best suit the application. This can be done on the basis of cross-validation. To illustrate this principle, the following section presents the application of the T-distributed error model to conventional X-ray data that are contaminated with outliers.

software d.f. Site	Careless ∞	Careless 16	Aimless ∞
C30-C115	22.38	25.60	22.28
C6-C127	17.84	16.77	16.33
C64-C80	23.34	24.21	21.64
C76-C94	19.62	19.93	19.44
M105	17.06	19.04	19.92
M12	18.39	20.34	20.50

Table 1: Anomalous peak heights for the sulfur atoms in the heavy atom substructure solution of hen egg white lysozyme from Autosol [13].

3 Results

3.1 Careless allows for robust inference of small crystallographic signals

To begin, we assessed the ability of Careless to extract small crystallographic signals from conventional monochromatic rotation series data.

To do so, we applied `careless` to a sulfur single-wavelength anomalous diffraction (SAD) data set of hen egg white lysozyme. It consists of a single 1,440 image rotation series in 0.5 degree increments collected at a low X-ray energy, 6.5 keV recorded at Advanced Photon Source beamline 24-ID-C. The refined structure is deposited in the Protein Data Bank (PDBID: 7L84), and the raw diffraction images will be made available through the SBGrid Data Bank. The reason these data are challenging to analyze is threefold. Firstly, they were collected at ambient temperature and very low flux in order to limit radiation damage. Approximately 50% of the pixels on each image did not record an event. Secondly, there is a shadow from the beam stop mounting bracket on each image near the edge of the detector in the 2 to 2.2 Å range. Finally, there was some leakage from a higher energy undulator harmonic which means there is 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 5a).

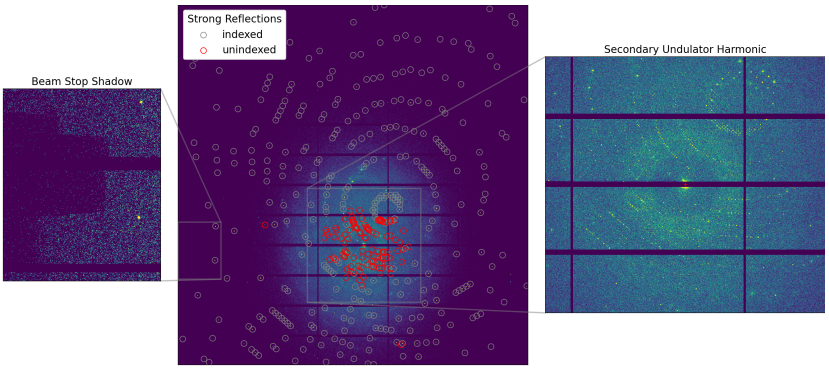
To address the outliers, we used cross-validation to select an appropriate degrees of freedom (d.f.) parameter of the likelihood for these data. In titrating this parameter, we found that a value of 16 degrees of freedom resulted in the largest $CC_{1/2}$ value (Figure 5b). Reassuringly, the region around 16 degrees of freedom exhibits comparatively high half data set anomalous correlations in the intermediate resolution range and suppressed spurious anomalous signal in the low and high resolution ranges attributable to outliers (Fig 5c, 5d). As is apparent in Figure 5d, the tuned likelihood function recovers comparable signal to the conventional merging program, Aimless[1]. However, with a normal likelihood (∞ d.f.) the model becomes fixated on the outliers in the high resolution region of reciprocal space.

In order to validate the anomalous correlations in real space, we used Autosol[13] to phase our merging results using the single wavelength anomalous diffraction method. Specifically, we compared the `careless` output with ∞ or 16 d.f. to the same data merged by the conventional method, Aimless. In order to ensure a consistent solution, 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* and automatically build a model (not shown). It is clear from the density-modified experimental maps, Figure 5e,5f,5g that the `careless` output with the normally distributed error model (∞ d.f.) is much lower quality. By contrast, both `careless` with 16 d.f. and aimless produced clearly interpretable experimental maps. The refined heavy atom structure factors yielded similar sulfur peak heights (Table 1) 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.

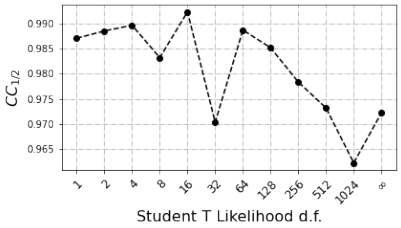
3.2 Polychromatic Data Merging with Careless

The global scaling model incorporated in `careless` is flexible with respect to the metadata it can leverage. One particularly intriguing use case, is in the reduction of polychromatic X-ray data. In this sort of experiment, a crystal is excited by an X-ray beam with a larger bandwidth than a conventional source. Usually referred to as Laue diffraction, these experiments contain reflections over a variety of wavelengths which has several benefits. Most importantly, enough photons can be delivered in a very short pulse in order to record a diffraction image allowing Laue crystallographers to

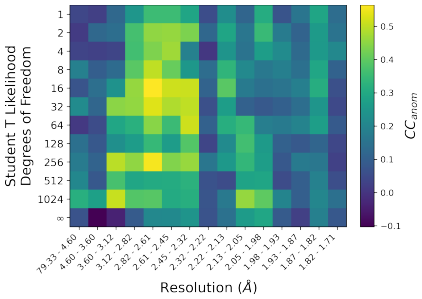
(a) Sources of outliers in the lysozyme data set



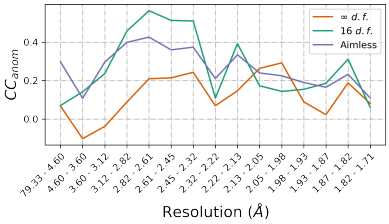
(b) Cross-validation of merging



(c) Anomalous half data set correlations



(d) Anomalous half data set correlations



(e) Careless (∞ d.f.) (f) Careless (16 d.f.) (g) Aimless

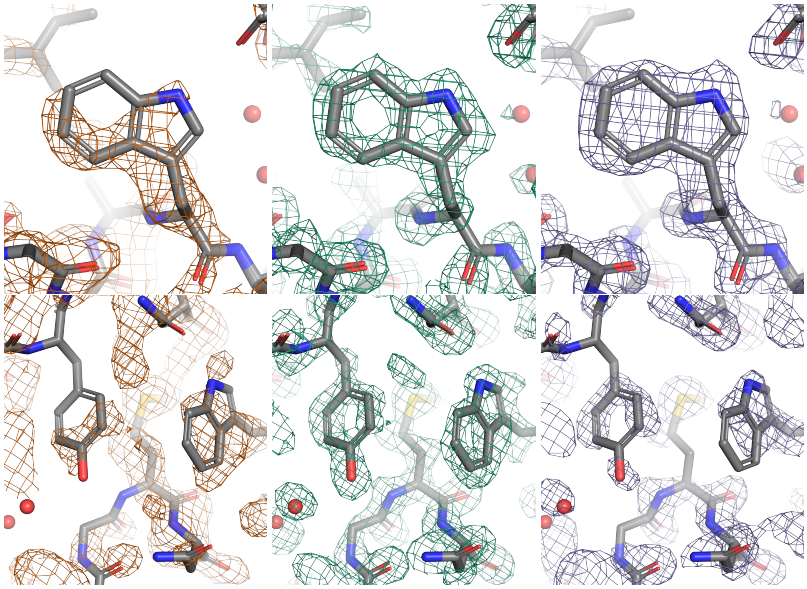


Figure 5: The results of processing the lysozyme sulfur SAD data set with `careless`. a) A sample diffraction pattern from the lysozyme data set indicating strong spots which could or could not be indexed by DIALS [12]. Insets show sources of outliers in the data. b) Cross-validation of `careless` merging as a function of the likelihood degrees of freedom. c-d) Resolution and d.f. dependent lysozyme anomalous signal. Panels e, f, g: Density modified experimental electron density maps produced with PHENIX Autosol [13] using the sulfur substructure from a reference structure (PDBID: 7L84) contoured at 1.5 σ. In these experiments, `careless` was minimized with the exact same parameters as the example on the GitHub page [10] except with 30,000 instead of 10,000 steps and except where the degrees of freedom has been changed to a value other than 16 or ∞.

probe time-dependent structural changes in the crystal on the picosecond timescale. Laue patterns, because of their polychromatic nature, also yield many more reflections per image than monochromatic still images making it a very attractive technology for serial crystallography.

Laue crystallography is, however, not without its drawbacks. The spectral nature of the 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. Complicating matters is the fact that reflections which lie on the same central ray in reciprocal space become superposed on the detector. A central ray is a ray in reciprocal space with its origin at $hkl = \{0, 0, 0\}$. These "harmonic" reflections need to be deconvolved algorithmically in order to be merged. Typical polychromatic data reduction software, 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 which is, ideally, proportional to the spectrum of the X-ray beam. After this is done, the contributions to each harmonic reflection can be deconvolved by solving a system of linear equations for each image. These complications make it a significant challenge to scale and merge polychromatic data. Perhaps it is not surprising that there are currently no open source merging packages supporting both wavelength normalization and harmonic deconvolution.

Fortunately, the forward modelling approach implemented in `careless` makes it relatively easy to account for these features of Laue diffraction. Firstly, in order to tackle wavelength normalization, all `careless` requires is access to the wavelength contributing to each reflection observation. Providing the empirical wavelength of each reflection observation from the experimental geometry enables the scale function to carry out wavelength normalization. In order 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

$$ELBO(\mu_{q_h}, \sigma_{q_h}, w) \approx \sum_{j=1}^{s} \left[\sum_{h,i} \log \mathcal{L} \left(I_{h,i} \middle| \sum_{r \in ray} F_{r,j}^2 \Sigma_{r,i,j}, \sigma_{I_{h,i}} \right) - \sum_{h} \left[\log q(F_{h,j}) - \log P(F_{h,j}) \right] \right], \quad (3)$$

which is readily optimized by same protocol demonstrated in Figure 4. 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.

To demonstrate that `careless` can effectively merge Laue data with high sensitivity, we applied it to a time resolved crystallography data set. The input data include 20 images from a single crystal of photoactive yellow protein (PYP) in the dark state and 20 images 2ms 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 (see Fig. 6a). These data were integrated in the Precognition (Renz Research, Inc.) Laue data reduction package. The unmerged reflections from Precognition were converted to `.mtz` files with a custom script based on `reciprocalspaceship`[14]. These files contain the intensity, uncertainty estimate, detector coordinates, and empirical wavelength of each reflection observation. As metadata, we supplied `careless` with the observed Miller index, wavelength, resolution, and detector position of each reflection observation. The dark and 2ms `.mtz` files were merged to produce separate outputs. We found that the quality of merging for this data set was quite high as judged by half data set correlation coefficients (Fig 6b, 6c). We then refined the dark data against a reference model (PDBID: 2PHY) [15]. We found 5 cycles of rigid body and isotropic B-factor refinement sufficient to generate high quality $2F_O - F_C$ electron density maps (Figure 6d). Using the phases from the 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 6e, these maps contained peaks in the vicinity of the PYP chromophore. In order to enhance the visualization of these maps, we applied a previously described weighting procedure [16]. The weighted maps (Figure 6f, 6g) show stronger difference density which is predominantly chromophore localized.

With previous generations of Laue merging software, it has been necessary to discard reflections below a particular I/σ 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. For the analysis presented here, we applied no such cutoff. Therefore, the ability of `careless` to identify the weak difference signals in this data set demonstrates an unprecedented degree of accuracy and a robustness to outliers. Furthermore, the appearance of interpretable difference electron density in the absence of a weighting scheme (Figure 6e) is equally exceptional. This observation strongly

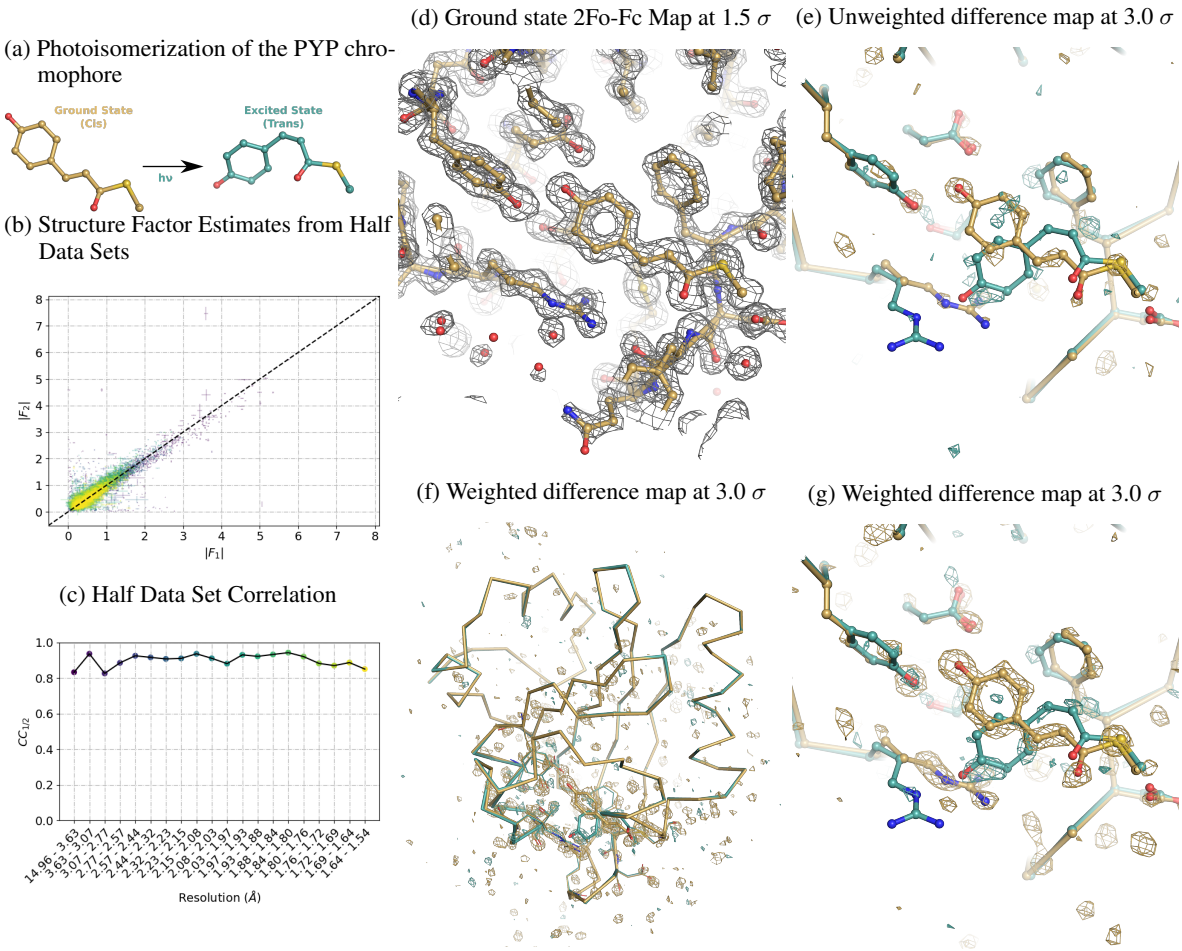


Figure 6: Merging time-resolved Laue crystallography data of photoactive yellow protein demonstrates that `careless` can merge polychromatic diffraction data robustly with high accuracy. (a) The PYP chromophore undergoes a trans to cis isomerization when exposed to blue light. In total, 40 images from a single crystal of PYP were processed. 20 were recorded in the dark state and 20 2ms after the arrival of a blue laser pulse. (b) Dividing the data in half and merging them with `careless` gave excellent correlation between the structure factor estimates. Error bars indicate the standard deviation of the structure factor posteriors for the half data sets. Points are colored by resolution as in (c) Plot of the half data set correlation coefficients color by resolution bin including both the dark and 2ms data. (d) Ground state 2Fo-Fc map created by refining the ground state model (PDBID: 2PHY) against the dark merging results (e) $F^{2ms} - F^{dark}$ time-resolved difference map showing the accumulation of blue positive density around the excited state chromophore (blue model, PDBID: 3UME) (f) $F^{2ms} - F^{dark}$ weighted time-resolved difference map showing localization of the difference density to the region surrounding the chromophore. (g) $F^{2ms} - F^{dark}$ weighted time-resolved difference map showing large differences around the chromophore.

suggests that `careless` represents a substantial improvement over the state of the art for the analysis of time-resolved Laue experiments

3.3 Merging X-ray Free-Electron Laser Data using Careless

To further establish generality of variational inference as a tool for merging X-ray data, we used `careless` to merge serial crystallography data from an X-ray free-electron laser experiment. To do so, we introduced an option to in the command line interface. The default merging model introduced in Figure 3 is most suited to merging single crystal data wherein the scale corrections vary smoothly across a rotation series. In serial crystallography, each image originates from a different sample with a different amount of scattering mass. A completely global scaling model is therefore not appropriate for such an experiment. In order to overcome this limitation, we exploited the modular design of the software to incorporate local image scale factors as shown in Figure 7a. This is a strategy used commonly in XFEL

and conventional merging programs [17, 18, 1, 19, 20]. With the inclusion of local scales, a_i, the augmented ELBO becomes

$$ELBO(\mu_{q_h}, \sigma_{q_h}, w, a) \approx \sum_{j=1}^{s} \left[\sum_{h,i} \log \mathcal{L}\left(I_{h,i} | a_i F_{h,j}^2 \Sigma_{h,i,j}, \sigma_{I_{h,i}}\right) - \sum_{h} \left[\log q(F_{h,j}) - \log P(F_{h,j})\right] \right], \quad (4)$$

where $a_1 = 1$ is fixed during optimization to avoid introducing equivalent classes of solutions.

The unmerged XFEL data we analyzed are available from the Coherent X-ray Imaging Data Bank (CXIDB) [21] under entry ID 81. In this experiment[22], a slurry of thermolysin microcrystals was delivered to the XFEL beam by a liquid jet. The data were recorded at an X-ray wavelength of 1.27 Å and contain significant anomalous signal from the zinc and calcium ions in the structure. The data were integrated[23] with DIALS[12], and the full analysis including all intermediate files is deposited in the CXIDB. For this experiment, we analyzed a single run containing 3,160 images. Using a custom script, the integrated intensities from CXIDB were converted to an mtz file. Both are available on the `careless` GitHub page [10]. This file contains as metadata the detector coordinates and Ewald offset of every reflection. For a still image, the maximal intensity for a given reflection is observed on the detector if the so-called Ewald sphere interesects the reflection centroid. The Ewald offset (EO) is a scalar value which summarizes the degree to which a particular reflection observation deviates from its maximum attainable diffraction intensity (Figure 7b). To test if our model could leverage this information about experimental geometry, we applied `careless` with local image scales to the unmerged intensities with and without the inclusion of the Ewald offset metadatum. The resulting merged structure factors were phased by isomorphous replacement from PDBID 2TLI [24]. With the EO, our protocol yielded markedly superior results as judged by the half dataset correlation coefficient (Figure 7d) as well as the refinement residuals (Figure 7e). To verify that including the metadata actionably improved the information content of the output, we constructed anomalous omit maps. The anomalously scattering ions in the reference structure were omitted during refinement against the `careless` output. The anomalous difference peak heights at the former location of each of the anomalous scatterers, tabulated in Figure 7c, strongly suggest the inclusion of Ewald offsets actionably improves the resulting structure factor estimates.

4 Discussion

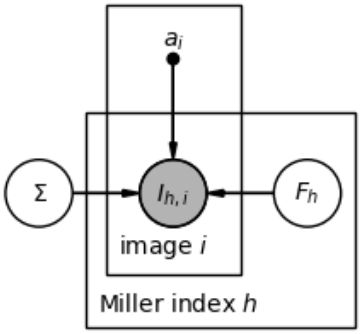
4.1 Physical Models Recapitulated in These Results

We have demonstrated in this manuscript that variational inference without explicit physical models can be very successful in scaling and merging X-ray diffraction data. This begs the question, what sorts of physical effects does variational inference account for in modeling these data? Unfortunately, our scale function operates in a high dimensional space making it difficult to interrogate. In our own experiments, we have found that inclusion of the reflection resolution is essential for proper data reduction in `careless`. This suggests that, one of the most basic corrections the model learns is an isotropic scale factor akin to the per image temperature factors which many merging packages include. 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 indicating that for some data sets, `careless` is able to learn an effective anisotropic scaling model. Likewise, the inclusion of the detector positions of the reflections seems to increase merging performance. Furthermore, it seems to be essential in cases where no polarization correction has been applied to the unmerged intensities such as the Laue and XFEL data presented here. This hints at the possibility that source polarization is implicitly corrected by the model. The success of merging Laue data would suggest that `careless` is able to infer the spectrum of the source as well or a projection thereof. Finally, the XFEL example suggests the model is competent to infer partialities in still images. Unlike other approaches [25, 18, 26] for partiality inference, `careless` does not postrefine the crystallographic orientation for each frame. Therefore, it must rely entirely on the accuracy of the geometric parameters inferred in upstream processing. It remains to be seen whether this approach will be valid in cases with poor refined geometry.

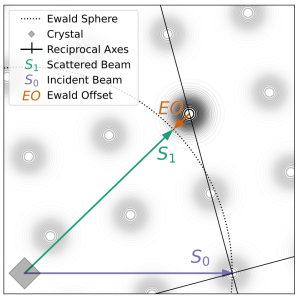
4.2 Scale Function Priors

At present, `careless` does not enforce any particular prior distribution on the scale factors, Σ. A thoughtful prior on this portion of the model would likely increase the robustness of the model by limiting overfitting. Given the variability in detector physics and sample quality it is hard to imagine a correct and general parameterization of a prior distribution over scale factors. However, it may be possible to develop a family of tunable prior distributions which could be tailored to the physics of a particular experimental apparatus. We imagine that in the near future, beamline scientists will be able to develop their own prior distributions bespoke for each experimental end station. Future work should focus on protocols for developing and optimizing such prior distributions.

(a) Augmented Graph with Maximum Likelihood Image Scales



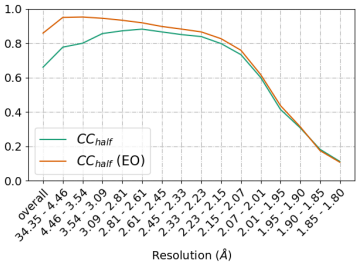
(b) Diffraction Condition for Still Images



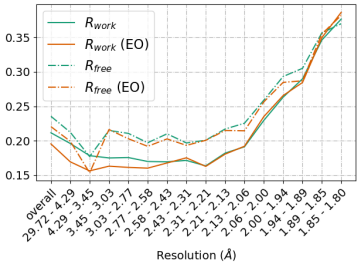
(c) Anomalous Peak Heights

Ewald Offset Atom	No	Yes
Ca-318	4.13	4.40
Ca-319	<3.0	3.76
Ca-320	4.14	5.85
Ca-321	4.34	4.96
Zn-317	11.66	15.90

(d) Half Data Set Correlation Coefficient



(e) Refinement R Value



(f)

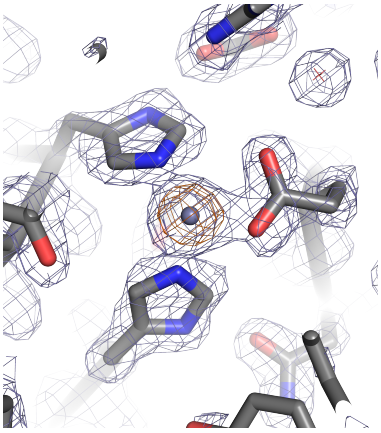


Figure 7: `careless` can leverage geometric metadata to improve XFEL data merging. a) By including local image scale parameters, `careless` can scale XFEL data. The parameters a_i can simply be allowed to optimize by maximum likelihood. b) The Ewald offset, shown in red, can optionally be included during refinement. c) Peak heights of anomalous scatterers in an anomalous omit maps, in σ units, for `careless` output with ("Yes") and without ("no") Ewald offsets. f) `careless` $2F_O - F_C$ electron density map contoured at 1.5 σ overlayed with thermolysin anomalous omit map contoured at 5.0 σ. Data merged with EO.

4.3 Stochastic Training

Although not addressed in this manuscript, it should be possible to implement stochastic training for the `careless` model. In stochastic training, gradient descent is conducted on a subset of the data at each iteration. This training paradigm, would allow variational inference to be employed in merging data sets too large to fit in memory on single accelerator card. If stochastic training can be successfully implemented for the `careless` model, it would be an ideal candidate for merging data at the next generation of X-ray free-electron lasers which expect to see data acquisition rates in the kilo to megahertz regime. Stochastic training could merge these data on the fly during data acquisition enabling quicker feedback during X-ray free-electron laser experiments.

5 Acknowledgements

We thank Vukica Srajer and Marius Schmidt for providing Laue time-resolved data, and Nick Sauter and Aaron Brewster for helpful discussions. Thanks to the staff at the Northeastern Collaborative Access Team (NE-CAT), beamline 24-ID-C of the Advanced Photon Source for supporting our room temperature crystallography experiments with special thanks to Igor Kourinov. NE-CAT beamlines supported by the NIH by the National Institute of General Medical Sciences from the National Institutes of Health (P30 GM124165). This research used 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." J.B.G. was supported by the National Science Foundation Graduate Research Fellowship under Grant No. DGE1745303.

6 Data Availability

The three data sets discussed in this manuscript have been adapted into examples available through the `careless` GitHub page [10] including the relevant merging and refinement scripts.

References

[1] 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.

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

[3] 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.

[4] 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.

[5] Rajesh Ranganath, Sean Gerrish, and David M. Blei. Black Box Variational Inference. *arXiv:1401.0118 [cs, stat]*, December 2013.

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

[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] 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.

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

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

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

[12] 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. *Acta Crystallographica Section D: Structural Biology*, 74(2), February 2018.

[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.

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

[15] 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.

[16] 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.

[17] Thomas A. White, Richard A. Kirian, Andrew V. Martin, Andrew Aquila, Karol Nass, Anton Barty, and Henry N. Chapman. *CrystFEL*: a software suite for snapshot serial crystallography. *Journal of Applied Crystallography*, 45(2):335–341, Apr 2012.

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

[19] Andrew G. W. Leslie. The integration of macromolecular diffraction data. *Acta Crystallographica Section D*, 62(1):48–57, 2006.

[20] Zbyszek Otwinowski and Wladek Minor. Processing of x-ray diffraction data collected in oscillation mode. In *Macromolecular Crystallography Part A*, volume 276 of *Methods in Enzymology*, pages 307–326. Academic Press, 1997.

[21] Filipe R. N. C. Maia. The Coherent X-ray Imaging Data Bank. *Nature Methods*, 9(9):854–855, September 2012.

[22] 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.

[23] A. S. Brewster, D. G. Waterman, J. M. Parkhurst, R. J. Gildea, I. D. Young, L. J. O'Riordan, J. Yano, G. Winter, G. Evans, and N. K. Sauter. Improving signal strength in serial crystallography with DIALS geometry refinement. *Acta Crystallographica Section D: Structural Biology*, 74(9):877–894, September 2018.

[24] 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.

[25] 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.

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