Online Inference of Structure Factor Amplitudes for Serial X-ray Crystallography

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

Abstract

Advances in X-ray techniques at Free Electron Laser and synchrotrons now enable the collection of diffraction snapshots from millions of micro crystals. These are often paired with physical or chemical perturbations to obtain movies of the response of proteins to chemical and physical stimuli [1]. Analysis of these data requires scalable algorithms. Distributed computing is one way to accomplish this as national labs may provide the necessary compute resources. However, a more accessible approach would be to construct algorithms which can operate on small batches of data on a single computer. The extreme case, an online algorithm, learns to process data by looking at one example at a time. Here we describe the successful implementation of one such algorithm for scaling and merging reflection intensities. The algorithm uses deep learning to scale reflection intensities while encouraging the merged structure factor estimates to follow a crystallographic prior distribution. The model is trained by gradient descent on a Bayesian objective function. We demonstrate that the model can estimate productive global parameter updates from single images. This approach has modest hardware requirements, can adapt on the fly as new data are acquired, and has the potential for transfer learning between data sets. The algorithm can be the heart of a flexible, scalable infrastructure that powers the next generation of diffraction experiments.

1 A Bayesian Model for Structure Factor Estimation

In crystallography, the X-ray beam reveals the Fourier transform of the sample's electron density. Each image is a slice through the transform. Due to the periodic nature of crystal lattices, their X-ray scattering patterns consist of discrete puncta ("reflections") interspersed with signal that is largely due to background scattering. Reflections are windows into to the Fourier transform of the unit cell—the repeating unit that builds up the crystal lattice. Each reflection corresponds to a specific spatial frequency vector within the unit cell. The X-ray flux scattered to the reflections is proportional to the square of the corresponding Fourier amplitude or "structure factor" as it is known in the scattering literature. The goal of X-ray crystallography is to estimate a complete set of structure factor amplitudes to the best attainable resolution allowed by the sample. Along with phases, which can be

Machine Learning for Structural Biology Workshop, NeurIPS 2022.

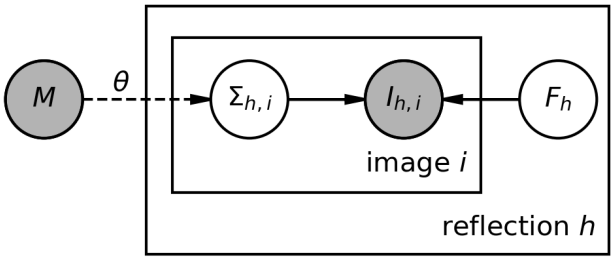


Figure 1: Probabilistic graphical model: For an image, i, the observed intensity, $I_{h,i}$, corresponding to a particular spatial frequency vector, h, depends on the structure factor F_h and a local random variable, $\Sigma_{h,i}$, representing the systematic error of the observation. As we showed previously [2], $\Sigma_{h,i}$ can be calculated as a function of per-observation metadata, M. In our previous study, this function takes the form of a multilayer perceptron with parameters θ.

recovered algorithmically or through specially designed experiments, the amplitudes determine the electron density of a sample.

Estimating structure factor amplitudes is complicated by copious sources of systematic error which need to be corrected. Many heuristic approaches exist to achieve these corrections. However, they ultimately rely on the physical intuition of the engineer to formulate suitable correction functions. As we showed in ref. [2], these errors can, instead, be corrected by a multilayer perceptron while simultaneously estimating structure factors using variational inference [3]. The approach does not require explicit physical modeling of error sources and recovers a set of structure factors satisfying two criteria: consistency with the observed data, and consistency with a prior distribution expressing first-principles statistical expectations.

In general, the intensity of reflections can be expressed as a graphical model (Figure 1). For this model, variational inference is accomplished by using gradient-based optimization to maximize the Evidence Lower BOund, [2]

$$\text{ELBO} = \sum_h \sum_i \left\{ \mathbb{E}_q \left[\log p(I_{h,i} | q_{F_h}, q_{\Sigma_{h,i}}, \sigma_{I_{h,i}}) \right] - D_{KL}(q_{\Sigma_{h,i}} \| p_\Sigma) \right\} - \sum_h D_{KL}(q_{F_h} \| p_F) \quad (1)$$

I and σ_I refer to observed reflections intensities and their error estimates. q_F and q_Σ are the variational distributions for structure factors and multiplicative scale factors with corresponding prior distributions p_F and p_Σ. The particular parameterization of this objective is a modeling choice. Our previous implementation relied on local parameters for each diffraction image in order to learn q_Σ which prevented batch training. Here we introduce a global parameterization of q_Σ that lifts this restriction enabling online inference (Figure 2).

2 Model Parameterization

As in our previous study [2], we parameterize the posterior of structure factors, q_F, by a truncated normal with positive support. For the prior, p_F, we use Wilson's priors, a historical prior from the crystallography literature [4]. We choose to parameterize the posterior of scales, q_Σ, as a lognormal distribution with a prior, $p_\Sigma = \mathcal{L}og\mathcal{N}ormal(0, 1)$. The parameters of q_Σ are amortized by a model discussed in the following sections. Using the fact that intensities are proportional to the square of the structure factor amplitudes, the likelihood

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

is taken to be a normally distributed about the product of the scales and the square of the structure factor amplitudes with the empirically measured standard deviation, $\sigma_{I_{h,i}}$.

For a particular observation with intensity $I_{h,i}$ and empirical uncertainty $\sigma_{I_{h,i}}$, we approximate the ELBO (Equation 1)

$$\begin{aligned} \text{ELBO}_{h,i} \approx \sum_{s=1}^{S} \Big\{ & \log \mathcal{N}(k I_{h,i} | \Sigma_{h,i,s} F_{h,s}^2, k\sigma_{I_{h,i}}) - w_F \left[\log q_{F_h}(F_{h,s}) - \log p_F(F_{h,s}) \right] \Big\} \\ & - w_\Sigma D_{KL}(q_{\Sigma_{h,i}} \| p_\Sigma) \end{aligned}$$

In this context, we use $F_{h,s}$ and $\Sigma_{h,i,s}$ to denote reparameterized samples [5],

$$F_{h,s} \sim q_{F_h}$$
$$\Sigma_{h,i,s} \sim q_{\Sigma_{h,i}}$$

from the variational distributions. The final term, which is the Kullback-Leibler divergence between the scale, q_Σ, and its its prior is computed analytically. The weight terms, w_F and w_Σ, are hyperparameters which modulate the strength of the priors. k is a strictly positive learned parameter which accounts for the arbitrary mean of p_Σ and is constrained by the softplus function. We note that there is a separate, global variational distribution, q_F, for each reflection, h. At each training step, the ELBO is summed over the reflection observations from a single of image, leading to the objective function,

$$\mathcal{L}_i = -\sum_h \text{ELBO}_{h,i}$$

which is minimized during training.

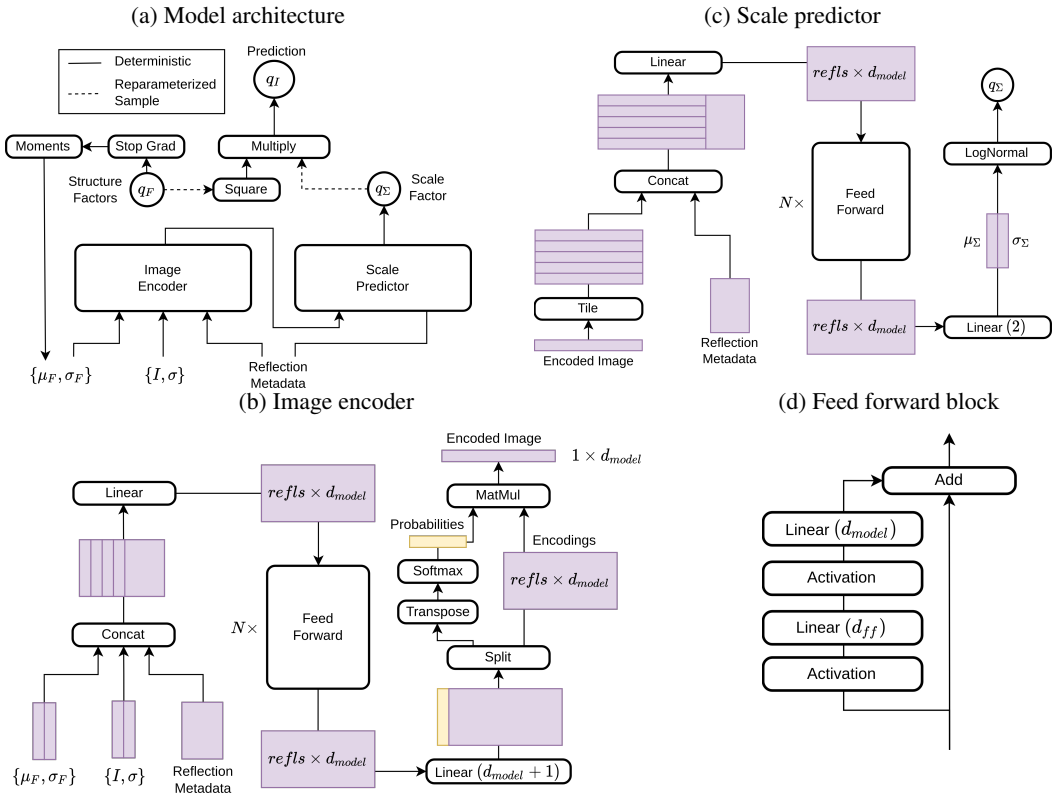


Figure 2: (a) The model predicts scaled intensities for a batch of reflections from one image. (b) The image encoder uses the current state of the structure factors, the intensities, and metadata for each reflection to compute a putative image encoding and corresponding probability. The output is the expectation of these putative encodings. (c) The reflection scale model concatenates encoded images to their corresponding reflection metadata. The concatenated vectors are passed through a neural network which outputs 2 parameters that the model interprets as the location and scale parameters of a lognormal distribution. (d) The basic nonlinear building block of the model is the ResNet v2 [6] without normalization layers. d_{ff}, d_{model}, and the choice of activation function are hyperparameters. Kernel weights are initialized by $W_{ij} \sim \mathcal{T}runcated\mathcal{N}ormal(0, \sigma, -2\sigma, 2\sigma)$ with $\sigma = \sqrt{2/[5N(d_{in} + d_{out})]}$. Biases are initialized to zero.

2.1 Amortized Inference for Reflection Scales

The scale model (Figure 2a) must learn to predict the variational distributions, $q_{\Sigma_{h,i}}$ from image metadata. In order to enable stochastic training, this must be accomplished by purely global parameters. To achieve this, we introduce a bipartite model (Figure 2a for image scales which consists of

a permutation invariant encoder model (Figure 2b) and a multilayer perceptron (Figure 2d). This architecture leverages the population of reflections within a given image to estimate appropriate scale parameters.

2.2 Image Encoder Model

The idea behind the encoder model (Figure 2b) is to learn a function which observes the current structure factor estimates, the context of each reflection (reflection metadata), and the observed intensities and suggests a vector that describes the image. Hopefully, this vector includes information needed to scale the reflections on this image. For instance, the vector should encode information about the orientation of the crystal, its size, shape, and mosaic properties, the brightness and deflection of the X-ray beam, etc. To generated this representation, the neural network considers each reflection in isolation and proposes a putative image encoding based on that reflection. Alongside the putative encoding, the network outputs a score representing the probability that the reflection is informative. The consensus encoding is taken to be the expected value of candidate encodings weighted by the probabilities. Pooling the candidate encodings into a single vector of length d_{model} prevents the encoder from passing the reflections' intensities directly to the scale model. This is essential to prevent the overall model from overfitting during training.

2.3 Scale Predictor Model

The goal of the scaling model is to produce a distribution of probable scale factors for each reflection (Figure 2c). As in previous work [2], it uses metadata associated with each reflection in order to infer scales. Unlike the previously reported model, it has access to the image representation produced by the upstream encoder model. The input layers of this model simply concatenate the image vector onto each reflection's metadata. A neural network uses this enriched representation to predict appropriate scales. In the parameterization presented here, the neural network predicts a two-vector interpreted as the location and scale parameters of a log-normal distribution. We use the softplus function to constrain the scale to be positive.

3 Application to Single Wavelength Anomalous Diffraction Data

As a case study for the suitability of this model for merging serial femtosecond crystallography data, we applied it to a publicly available dataset consisting of 166,250 diffraction images [7] of the zinc metalloprotease, thermolysin, which were integrated using DIALS [8] (available from https://www.cxidb.org/ entry 81 under a public domain, CC0 license). The data were acquired at a wavelength of 1.27 Å. At this wavelength, several of the atoms in the thermolysin crystal absorb X-rays leading to a phenomenon known as "anomalous diffraction". Typically, the two centrosymmetrically related halves of the diffraction pattern have identical intensities. In the case of anomalous diffraction, this symmetry is broken. Small differences in intensities carry information about the phase of the structure factors which can be used to solve the structure of the sample. The strongest signal in this case is from the catalytic Zn^{2+} ions in the thermolysin active site. However, a number of calcium ions also contribute to the signal.

3.1 Training Details

We trained our model on the thermolysin data for 500,000 gradient updates using the Adam optimizer [9], equating to about 3 passes over the dataset. With the reported hyperparameters in Figure 3a, the model requires 0.8 GB of GPU memory and executes a gradient step in about 70ms on a consumer grade GPU (NVIDIA RTX 3060). The total training time was less than 12 hours. We expect a 3 to 10 fold speedup is attainable by optimizing the i/o pipeline.

3.2 Phasing Results

We noted the appearance of anomalous signal early in training as judged by the signal to noise of anomalous peaks corresponding to elements in the sample (Figure 4a). After 500,000 optimization steps, we analyzed the output using AutoSol [11], an automated package for estimating experimental phases from X-ray diffraction data. Our results were encouraging, surpassing the current state of

(a) Model hyperparameters

Metadata	Scattered wavevector x-coord (Å^{-1}), Scattered wavevector y-coord (Å^{-1}), Ewald offset vector (Å^{-1})
Activation	$ReLU$
N	20
d_{model}	32
d_{ff}	64
w_F	0.001
w_Σ	10
MC samples	50
Max refls per image	2048
Learning rate	0.001 $\times 50k$ 0.0001 after
β_1	0.9
β_2	0.999

(b) Training trajectory

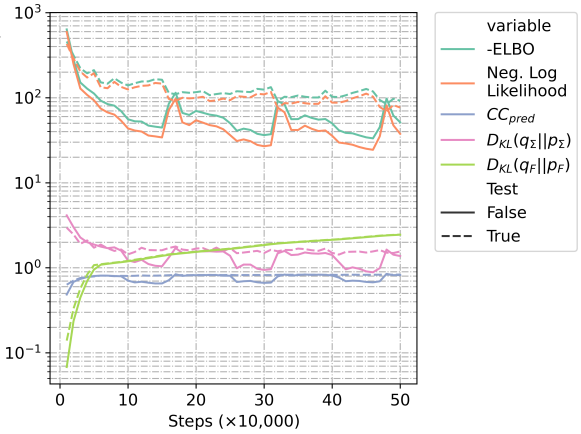
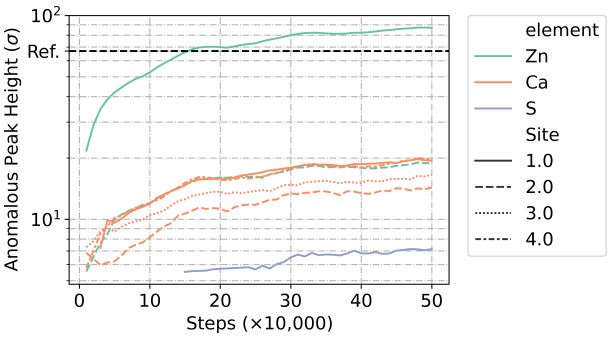


Figure 3: (a) The hyperparameters used in this thermolysin case study. (b) The trajectory of various metrics throughout training. This plot shows the -ELBO and its constituent terms, the log likelihood and Kullback-Leibler divergences. The CC_{pred} is the Spearman correlation between predicted and observed intensities calculated per image and averaged. The crossvalidation test set is 10% of the images.

(a) Anomalous signal during training



(b) SAD phasing results

	Ours	Reference [10]
Zn^{2+} peak (σ)	**83.2**	67.0
Model-map CC	**0.84**	0.80±0.001
Sites found	**7**	6
Residues built	**305**	297.2±6.5
R_{work}	**0.204**	0.212±0.013
R_{free}	**0.231**	0.237±0.017

Figure 4: (a) The development of anomalous signal throughout the cours of training. Anomalous peak heights determined using the difference map peak finding command line script from rs-booster (https://github.com/rs-station/rs-booster). (b) The table presents SAD Phasing Results from AutoSol [11] using the same parameters as [10] (bold is better). The zinc peak height in the table is from `phenix.find_peaks_holes` using our structure factor estimates and the autobuilt model.

the art 4b by a large margin. Specifically, we improved on the anomalous peak signal-to-noise by over 15 σ. As in the previous state of the art[8], we were able to automatically build a model of the thermolysin indicating that our inferred structure factors are precise and scaled such that they can be used in conventional processing software. Supporting the validity of our strategy, the experimentally determined electron density is readily interpretable immediately upon phasing and density modification (Figure 5a).

During our analysis, we discovered anomalous signal in a previously unreported site. The sulfur atom of methionine-205 appeared as an anomalous site after a single pass through the data (Figure 4a). Again, this signal was usable as judged by the fact that AutoSol was able to place a seventh anomalous atom during phasing (Figure 5b). This result is remarkable given that the available anomalous signal for sulfur at this wavelength is less than a single electron.

(a) Thermolysin active site (b) Methionine-205

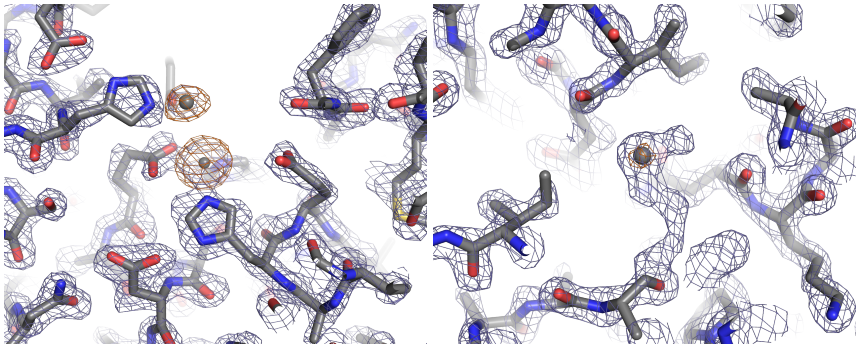


Figure 5: Thermolysin SAD phasing results featuring the autobuilt model from AutoSol [11]. The blue electron density, contoured at 2σ, is the experimental, density-modified map. The orange map is the anomalous difference map using the pictured model for phases and the structure factor estimates from our merging algorithm contoured at 5σ.

4 Conclusion

In this work we have demonstrated an algorithm which can estimate productive global updates for structure factor amplitudes given access to a single image at a time. We show a proof of concept result based on publicly available data and demonstrate performance surpassing the state of the art. As of yet, is it unclear how our choice of priors and hyperparameters will generalize.

Our model requires relatively meager resources to train. The memory requirements scale linearly with the number of reflections per image. Therefore, we submit it may be possible to apply this model to raw diffraction images which are typically in the 10 megapixel regime. This would allow us to sidestep the current, error-prone practice of independently estimating the flux to each reflection observation. We imagine the only pre-processing which will be required in this case is to assign each pixel to its nearest structure factor, a task that can be accomplished with existing software.

Acknowledgments and Disclosure of Funding

We would like to thank Aaron Brewster, Derek Mendez, and Jack Greisman for many helpful conversations about crystallography. We are grateful to Minhuan Li, Ian Hunt-Isaak, and John Russell for input on machine learning algorithms and libraries. This work was supported by the Searle Scholarship Program (SSP-2018-3240), a fellowship from the George W. Merck Fund of the New York Community Trust (338034), and the NIH Director's New Innovator Award (DP2-GM141000). KD holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund.

References

[1] 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. 9(1):4025. Number: 1 Publisher: Nature Publishing Group.

[2] Kevin M. Dalton, Jack B. Greisman, and Doeke R. Hekstra. Careless: A variational bayesian model for merging x-ray diffraction data. page 2021.01.05.425510. Publisher: Cold Spring Harbor Laboratory Section: New Results.

[3] David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. Variational inference: A review for statisticians. 112(518):859–877.

[4] A. J. C. Wilson. The probability distribution of x-ray intensities. 2(5):318–321. Number: 5 Publisher: International Union of Crystallography.

[5] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes.

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. Number: arXiv:1603.05027.

[7] 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. 5(1):4371. Number: 1 Publisher: Nature Publishing Group.

[8] 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. 74(9):877–894. Number: 9 Publisher: International Union of Crystallography.

[9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization.

[10] A. S. Brewster, A. Bhowmick, R. Bolotovsky, D. Mendez, P. H. Zwart, and N. K. Sauter. SAD phasing of XFEL data depends critically on the error model. 75(11):959–968. Publisher: International Union of Crystallography.

[11] 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. 65(6):582–601. Number: 6 Publisher: International Union of Crystallography.