Integrating Macromolecular X-ray Diffraction Data with Variational Inference

Luis A. Aldama
The Harvard Biophysics Graduate Program
Harvard Medical School Campus
240 Longwood Avenue, Seeley G. Mudd Bldg., Room 204c
Boston, MA 02115
luis_aldama@g.harvard.edu

Kevin M. Dalton
Linac Coherent Light Source, SLAC National Accelerator Laboratory, Menlo Park, 94025, CA, USA
Department of Biology, New York University, New York, NY 10012
Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA 02138
kmdalton@slac.stanford.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

X-ray crystallography is a fundamental technique for determining the three-dimensional structure of biological macromolecules. The method involves crystallizing the molecule, exposing it to X-ray beams, and recording the resulting diffraction patterns, which consist of discrete signals called reflections. A critical computational step is the integration of diffraction data, where reflection intensities must be accurately estimated from noisy images.

We propose a deep learning-based integration algorithm that leverages variational inference to infer the intensity and background of each reflection under a Poisson-likelihood model. Preliminary results show that our model's integrated intensities underperform on crystallographic metrics compared to state-of-the-art algorithms. However, our integrated intensities were successfully used to reconstruct an electron density map and atomic model.

1 Introduction

X-ray crystallography provides the theoretical and experimental framework for determining the electron density of a molecule-of-interest. In structural biology, X-ray crystallography is used to construct three-dimensional electron density maps and atomic models of biological macromolecules, such as proteins and nucleic acids [1]. The resulting atomic models, typically deposited to the Protein Data Bank [2], can be leveraged to design more efficient therapeutics [3, 4, 5], in molecular dynamics simulations, and as training data for tasks such as structure prediction and drug design [6, 7].

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

An X-ray crystallography experiment involves crystallizing a molecule, passing an X-ray beam through the crystal, and capturing the resulting diffraction patterns on a detector. The diffraction patterns contain *reflections* which are related to the molecule's structure [8]. The patterns are recorded in 2-dimensional arrays which are stacked into a rank-three tensor of voxels. The location of each reflection within the voxel grid is dependent on the geometry of the crystal lattice, whereas the intensity of each reflection contains structural information about the molecule. Specifically, the intensity is proportional to the square of the Fourier amplitudes of the electron density (Figure 1). The Fourier transform of the electron density are known as a *structure factors*. Estimating the structure factor amplitudes from the diffraction patterns is a key step in crystallographic data analysis. In order to do so, reflection intensities must be estimated from noisy diffraction data with an algorithm called *integration* [9].

We propose a deep-learning-based algorithm for integrating X-ray diffraction data using variational inference [10, 11]. Our method models photon counts at each pixel using a Poisson distribution, treating the true intensity and background as latent variables. By leveraging neural network architectures, we aim to provide a flexible alternative to traditional integration methods. We show that our model can be used to integrate macromolecular X-ray diffraction data, and that the results can be used to successfully determine the electron density of an experimental dataset.

2 Background: From Crystal to Structure

This section provides a brief description of the experimental and computational pipelines used in X-ray crystallography. Given a molecule-of-interest, the ultimate goal of a crystallography experiment is to determine its electron density, described by a function $\rho(\mathbf{r}) : \mathbb{R}^3 \to \mathbb{R}$ that maps a position in the crystal $\mathbf{r}$ to a real-valued amount of electron density. This is achieved through a combination of experimental and computational work. In the lab, the molecule-of-interest is expressed, purified, and crystallized [1, 12]. These crystals are used in *diffraction experiments* to collect *diffraction patterns* (Figure 1). An X-ray beam is passed through the crystallized sample, and a fraction of the X-ray beam interacts with the electrons in the sample. At specific angles of interaction, constructive interference occurs between the diffracted X-rays, producing distinct signals known as *reflections* or *spots* on the detector [8]. The geometry of the crystal lattice, rather than its contents, determines where constructive interference will occur.

After the experiment, the diffraction images are processed through a series of algorithms known as *data reduction* [13, 14, 15, 16]. Data reduction consists of (1) *spot finding & indexing* where the Mlller index $\mathbf{h}$ of each reflection is determined, (2) *integration* where the intensity $I_{\mathbf{h}} \propto |\mathbf{F}_{\mathbf{h}}|^2$ of each observed reflection is estimated, and (3) *scaling and merging* where the scaled and merged structure factor amplitude for each Miller index $|\mathbf{F}_{\mathbf{h}}|$ is calculated. The phase of each Miller index, $\varphi_{\mathbf{h}}$, can be determined with experimental phasing techniques, or through molecular replacement [17, 18]. Together, the Miller index $\mathbf{h}$, the scaled and merged amplitude $|\mathbf{F}_{\mathbf{h}}|$, and the phase $\varphi_{\mathbf{h}}$, are called the *structure factor*, denoted by $\mathbf{F}_{\mathbf{h}} = |\mathbf{F}_{\mathbf{h}}|e^{(i\varphi_{\mathbf{h}})}$. The electron density is the inverse Fourier transform of the structure factors.

Accurate electron density estimation requires precise determination of each component of the structure factor [19, 20]. Traditional integration methods, such as direct summation and profile fitting, face challenges with weak signals and noisy data. Profile fitting, for instance, relies on predefined *standard profiles*—models that describe the expected intensity distribution across the detector for strong reflections. While effective in many cases, these profiles are based on idealized assumptions and can introduce systematic errors when experimental data deviates from them [21, 9].

Advancements in X-ray sources, detectors, and sample delivery techniques have shifted focus towards capturing molecular dynamics [22, 23, 24]. These developments introduce new challenges in signal extraction from increasingly complex diffraction data [20, 15, 24]. Our work explores the application of deep learning algorithms to macromolecular X-ray diffraction data processing, aiming to address these challenges. We propose a Bayesian model that does not rely on a physical model of the diffraction experiment, potentially allowing for generalization across different experimental modes.

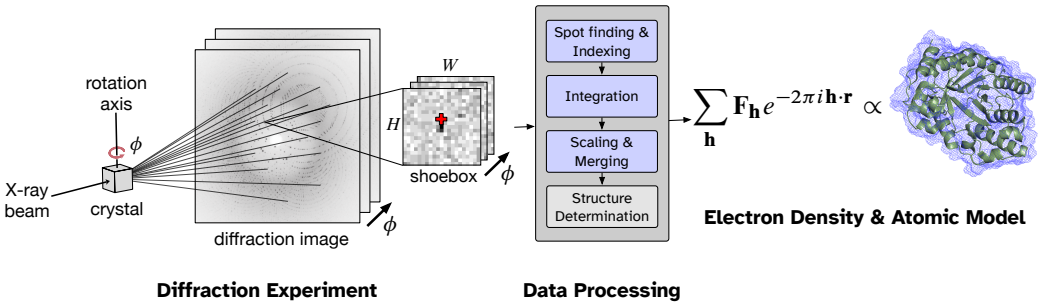

Figure 1: Diagram of an X-ray crystallography experiment. Diffraction images record the diffracted intensities from an X-ray passing through a crystalline sample. The diffracted X-rays resemble ellipsoids, and they are called *spots* or *reflections*. The intensity, $I_{\mathbf{h}} \propto |\mathbf{F}_{\mathbf{h}}|^2$, and phases, $\varphi_{\mathbf{h}}$ are determined using *data reduction* and phasing techniques. These estimates are used to find the inverse Fourier transform of the structure factors $\mathbf{F}_{\mathbf{h}}$.

3 Model Parameterization

Given a dataset of N reflections, our goal is to estimate the signal intensity and background scattering of each reflection, denoted by I_i and Bg_i for $i \in \{0, ..., N\}$, respectively. Each reflection, denoted by $\mathbf{X}_i$, is represented as a three-dimensional array of voxels called a *shoebox*. Each voxel in the shoebox contains a value x_{ij} that represents the number of photons counted during X-ray exposure.

We model the photon counts x_{ij} at each pixel as being drawn from a Poisson distribution with rate parameter λ_{ij}: $x_{ij} \sim \mathrm{Poisson}(\lambda_{ij})$, where the rate parameter λ_{ij} is a function of the reflection intensity I_i, background Bg_i, and the reflection's profile weight γ_{ij}:

$$\lambda_{ij} = I_i \times \gamma_{ij} + \mathrm{Bg}_i. \tag{1}$$

Here, I_i and Bg_i are latent variables, meaning they are not directly observed but must be inferred from the data. To infer these variables, we employ a Bayesian framework. We encode the dependencies between the observed data and the latent variables with a graphical model (Figure 2) [25]. Our goal is to compute the posterior distribution $p(I_i, \mathrm{Bg}_i, \gamma_{ij}|x_{ij})$. Using Bayes' theorem,

$$p(I_i, \mathrm{Bg}_i, \gamma_{ij}|x_{ij}) \propto p(x_{ij}|I_i, \mathrm{Bg}_i, \gamma_{ij})p(I_i)p(\mathrm{Bg}_i)p(\gamma_{ij}), \tag{2}$$

where $p(xi_i|I_i, \mathrm{Bg}_i, \gamma_{ij})$ is the likelihood, and $p(I_i)$, $p(\mathrm{Bg}_i)$, and $p(\gamma_{ij})$ are the prior distributions for the intensity, background, and profile, respectively.

To approximate the posterior, we use Variational Inference (VI), which allows us to optimize a family of parameterized distributions $\mathcal{Q}$. The goal is to find the distribution $q(\mathbf{z}) \in \mathcal{Q}$,where $\mathbf{z} = (I_i, \mathrm{Bg}_i, \gamma_{ij})$, that maximizes the Evidence Lower Bound (ELBO):

$$\mathrm{ELBO}(q) = \mathbb{E}_{q(\mathbf{z})}\left(\log P(\mathbf{x}|\mathbf{z})\right) - D_{\mathrm{KL}}(q(\mathbf{z})||p(\mathbf{z})), \tag{3}$$

where the first term is the expected log-likelihood, and the second term is the Kullback-Leibler (KL) divergence between the variational distribution $q(\mathbf{z})$ and the prior $p(\mathbf{z})$. The likelihood term favors solutions that best explain the data. The KL-divergence term penalizes solutions that deviate far away from our prior knowledge.

4 Model Architecture

Our model uses an autoencoder architecture, and it is composed of three main modules: an encoder, a parameter estimation module, and a decoder. All modules are written in PyTorch.

4.1 Encoder

The encoders are modified residual networks [26], and they encode shoeboxes into 64-dimensional latent vector representations (Figure 3). The first encoder uses fully connected (FC) layers, and the

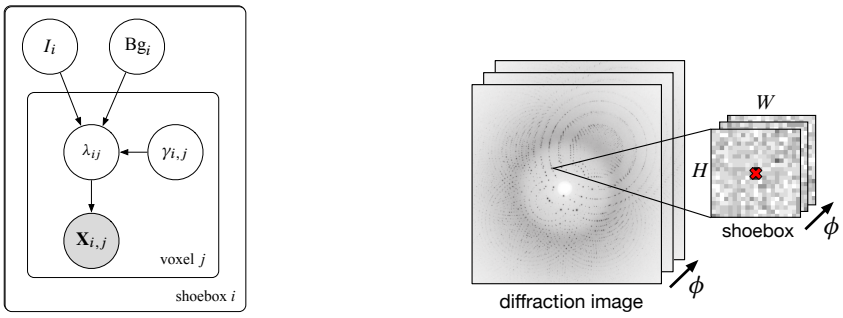

Figure 2: (left) Graphical model encodes the dependence of the observed counts x_{ij} on the latent variable λ_{ij}. (b) Diagram depicting a shoebox, which is a three-dimensional array of voxels extracted from a set of images. In practice, shoeboxes are stored as PyTorch tensors with shape (ϕ, H, W), where ϕ is equal to the number of images, and H and W are the height and width in number of voxels.

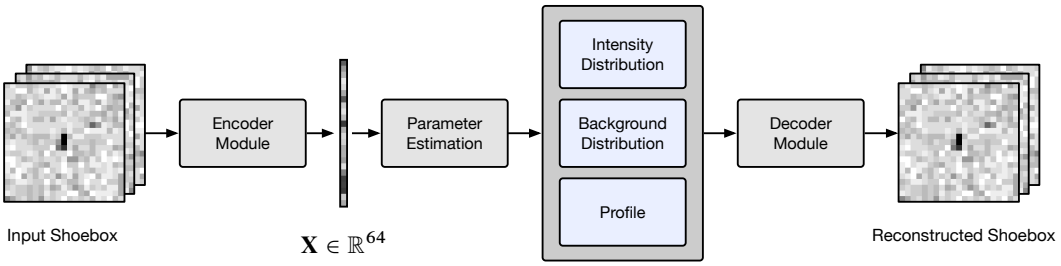

Figure 3: Diagram of the integration model architecture. The encoder takes aoebox as input and outputs a 64-dimensional vector. The parameter estimation module outputs variational distributions of the intensity and background, and it outputs the shoeboxes profile. These components are then passed through the decoder to reconstruct the input.

second uses convolutional (CN) layers. We refer to these architectures as the FC encoder and the CNN encoder, respectively.

The structure of the input data depends on the type of model that is used. In the FC encoder, each input shoebox is flattened into a vector $\mathbf{X} \in \mathbb{R}^{N_p \times 7}$, where N_p is the number of pixels, and seven corresponds to pixel-specific features (intensity, detector coordinates, and distance to centroid which is calculated upstream by DIALS by a least-squares fit to the experimental geometry). This vector passes through residual blocks, which output the latent representation vector $\mathbf{X} \in \mathbb{R}^{64}$.

In the CNN encoder, the shoebox data is preserved in its original spatial format, represented as $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$, where C represents the channels (pixel values and pixel-specific features), and $H \times W$ are the height and width of the shoebox.

4.2 Parameter Estimation Module

For each shoebox, the parameter estimation module generates the variational distributions of the intensity I_i and background Bg_i, and the profile γ_{ij} of each voxel. The intensity and background are modeled using two-parameter Gamma distributions, ensuring non-negative values. Additionally, the module estimates the profile of each reflection, which describes how the intensity is distributed across the shoebox. Profiles can be any integral 1 function $f(i, j) = \gamma_{ij} : \mathbb{R}^3 \mapsto \mathbb{R}_0^+$. We are exploring two profile models: (1) a Multivariate Normal (MVN) distribution, and (2) $\mathrm{Softmax}$-based profile. In either case, the profile acts as weight for each voxel, denoted by γ_{ij}, that scales the predicted intensity at each voxel (equation 1).

4.3 Decoder

The decoder uses the estimated parameters to generate reparameterized samples of the photon arrival rate λ_{ij} at each voxel. Intensity and background values are sampled from their respective variational distributions, denoted by $\mathbf{z}_I$ and $\mathbf{z}_{\mathrm{Bg}}$, and the rate is calculated as: $\lambda_{ij} = \mathbf{z}_{I_i} \times \gamma_{ij} + \mathbf{z}_{\mathrm{Bg}_i}$.

The model's objective is to maximize the ELBO, which is equivalent to minimizing the following loss function,

$$\mathrm{Loss} = -\mathbb{E}_q \left(\log p(x_{ij} \mid \lambda_{ij})\right) + \kappa_1 D_{\mathrm{KL}}(q(I) \parallel p(I)) + \kappa_2 D_{\mathrm{KL}}(q(\mathrm{Bg}) \parallel p(\mathrm{Bg})), \tag{4}$$

where $p(I)$ and $p(\mathrm{Bg})$ are the prior distributions of the intensity and background, respectively, and κ_1 and κ_2 are hyperparameters that control the strength of the KL divergence terms.

We use a Poisson distribution for the log-likelihood, which arises from the assumption that photon counts follow a Poisson process. For both the prior and variational distributions of intensity and background, we use Gamma distributions. Our choice of variational distributions is primarily driven by the positive support of the Gamma distribution, which aligns with the physical reality that intensities and background values cannot be negative.

5 Training and Preliminary Results

We validated our model using both simulated and real diffraction data. The experimental dataset consists of NaI-soaked hen egg-white lysozyme (HEWL) crystals, collected for anomalous scattering at the NE-CAT beamline 24-ID-C (Advanced Photon Source, Argonne National Laboratory). The dataset, recorded at 295 K using 11.95 keV X-rays, includes 1440 images captured on a PILATUS 6M-F detector (PDB ID: 9B7C) [27].

To assess model performance, we identified several metrics used in crystallographic data processing [13]. These metrics include $\mathrm{CC}_{1/2}$ (correlation between random half-datasets), $\mathrm{CC}_{\mathrm{anom}}$ (correlation of anomalous differences between half-datasets), R-factors (agreement between model and observed data), and anomalous peak counts and heights (which indicate the strength of anomalous signals) [13, 28]. These metrics enable comparison with existing state-of-the-art integration algorithms.

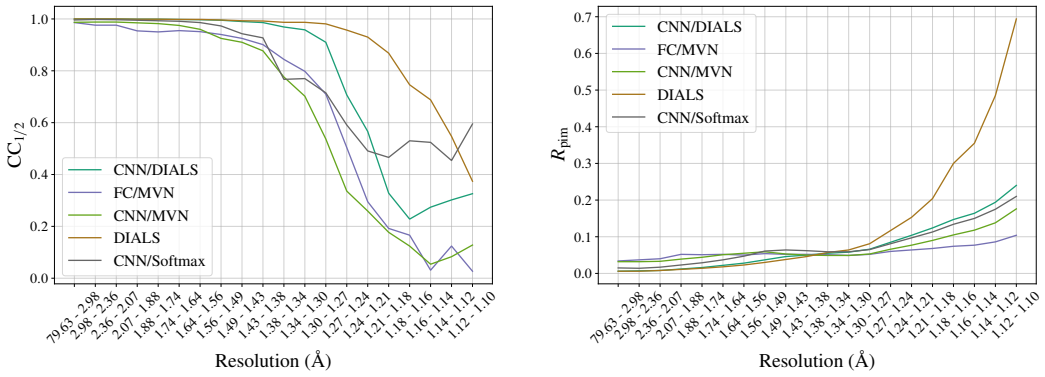

Figure 4: Figure of (left) $\mathrm{CC}_{1/2}$ as function of resolution (higher is better) and (right) R_{pim} as a function of resolution (lower is better).

We used three different model/encoder combinations to integrate the HEWL dataset: (1) CNN/MVN, (2) CNN/Softmax, and (3) FC/MVN. The same hyperparameters were used across all model combinations (Table 1. The resulting integrated intensities were then scaled and merged using `dials.scale` and `dials.merge` [29]. The scaled and merged structure factor amplitudes were then used with `phenix.refine` to build an electron density map and atomic model. Figure 4 and Tables 1 - 3 report the metrics obtained from our integration algorithm, and the DIALS integration algorithm.

6 Conclusion

We have demonstrated that deep learning and variational inference can be effectively applied to the integration of macromolecular X-ray diffraction data. Our evaluation pipeline includes metrics used in crystallographic data processing that allow us to evaluate the performance of our model. Preliminary results show that our integrated intensities underperform compared to the integration algorithms used in DIALS. However, our model's integrated intensities can be successfully used with downstream algorithms to reconstruct the electron density map and atomic model the corresponding molecule.

Our future work will focus on increasing the performance of the model. We will do so by exploring different choice of prior and variational distributions, extensive hyperparameter turning, and exploring alternative architectures such as Vision Transformers [30].

7 Acknowledgements

We would like to thank Graeme Winter for contributing to the code that converts diffraction images into PyTorch tensors. This work was supported by an NSF GRFP award to L.A.A (DGE 2140743), and an NIH457 Director's New Innovator Award (DP2-GM141000) to D.R.H. K.M.D is currently employed by the Linac Coherent Light Source (LCLS) at the SLAC National Accelerator Laboratory, an Office of Science User Facility operated for the U.S. Department of Energy Office of Science by Stanford University. K.M.D is supported by the U.S. Department of Energy, Office of Science, Basic Energy Sciences under Contract No. DEAC02-76SF00515.

References

[1] Bernhard Rupp. *Biomolecular Crystallography: Principles, Practice, and Application to Structural Biology*. Garland Science, New York, 1st edition edition, October 2009.

[2] F. C. Bernstein, T. F. Koetzle, G. J. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The Protein Data Bank: a computer-based archival file for macromolecular structures. *Journal of Molecular Biology*, 112(3):535–542, May 1977.

[3] Evanthia Lionta, George Spyrou, Demetrios Vassilatis, and Zoe Cournia. Structure-Based Virtual Screening for Drug Discovery: Principles, Applications and Recent Advances. *Current Topics in Medicinal Chemistry*, 14(16):1923–1938, October 2014.

[4] Leonardo Ferreira, Ricardo dos Santos, Glaucius Oliva, and Adriano Andricopulo. Molecular Docking and Structure-Based Drug Design Strategies. *Molecules*, 20(7):13384–13421, July 2015.

[5] Miles Congreve, Chris de Graaf, Nigel A. Swain, and Christopher G. Tate. Impact of GPCR Structures on Drug Discovery. *Cell*, 181(1):81–91, April 2020.

[6] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, August 2021.

[7] Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M. Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Shiyang Chen, Minjia Zhang, Conglong Li, Shuaiwen Leon Song, Yuxiong He, Peter K. Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21(8):1514–1524, August 2024. Publisher: Nature Publishing Group.

[8] C.G. Darwin. XCII. The reflexion of X-rays from imperfect crystals. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 43(257):800–829, May 1922. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/14786442208633940.

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

[10] 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.

[11] 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.

[12] Juan C. Sanchez, Melissa Carrillo, Suraj Pandey, Moraima Noda, Luis Aldama, Denisse Feliz, Elin Claesson, Weixiao Yuan Wahlgren, Gregory Tracy, Phu Duong, Angela C. Nugent, Andrew Field, Vukica Šrajer, Christopher Kupitz, So Iwata, Eriko Nango, Rie Tanaka, Tomoyuki Tanaka, Luo Fangjia, Kensuke Tono, Shigeki Owada, Sebastian Westenhoff, Marius Schmidt, and Emina A. Stojković. High-resolution crystal structures of a myxobacterial phytochrome at cryo and room temperatures. *Structural Dynamics (Melville, N.Y.)*, 6(5):054701, September 2019.

[13] Philip R. Evans. An introduction to data reduction: Space-group determination, scaling and intensity statistics. *Acta Crystallographica Section D Biological Crystallography*, 67(4):282–292, April 2011.

[14] 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 Structural Biology*, 76(4):385–399, April 2020.

[15] Kevin M. Dalton, Jack B. Greisman, and Doeke R. Hekstra. A unifying Bayesian framework for merging X-ray diffraction data. *Nature Communications*, 13(1):7764, December 2022.

[16] Luis A. Aldama, Kevin M. Dalton, and Doeke R. Hekstra. Correcting systematic errors in diffraction data with modern scaling algorithms. *Acta Crystallographica Section D Structural Biology*, 79(9):796–805, September 2023.

[17] Philip Evans and Airlie McCoy. An introduction to molecular replacement. *Acta Crystallographica Section D: Biological Crystallography*, 64(Pt 1):1–10, January 2008.

[18] Paul D. Adams, Pavel V. Afonine, Gábor Bunkóczi, Vincent B. Chen, Ian W. Davis, Nathaniel Echols, Jeffrey J. Headd, Li-Wei Hung, Gary J. Kapral, Ralf W. Grosse-Kunstleve, Airlie J. McCoy, Nigel W. Moriarty, Robert Oeffner, Randy J. Read, David C. Richardson, Jane S. Richardson, Thomas C. Terwilliger, and Peter H. Zwart. *PHENIX* : A comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography*, 66(2):213–221, February 2010.

[19] James M. Holton, Scott Classen, Kenneth A. Frankel, and John A. Tainer. The R-factor gap in macromolecular crystallography: An untapped potential for insights on accurate structures. *The FEBS Journal*, 281(18):4046–4060, September 2014.

[20] Derek Mendez, Robert Bolotovsky, Asmit Bhowmick, Aaron S. Brewster, Jan Kern, Junko Yano, James M. Holton, and Nicholas K. Sauter. Beyond integration: Modeling every pixel to obtain better structure factors from stills. *IUCrJ*, 7(6):1151–1167, November 2020.

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

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

[23] Melissa Carrillo, Suraj Pandey, Juan Sanchez, Moraima Noda, Ishwor Poudyal, Luis Aldama, Tek Narsingh Malla, Elin Claesson, Weixiao Yuan Wahlgren, Denisse Feliz, Vukica Šrajer, Michał Maj, Leticia Castillon, So Iwata, Eriko Nango, Rie Tanaka, Tomoyuki Tanaka, Luo Fangjia, Kensuke Tono, Shigeki Owada, Sebastian Westenhoff, Emina A. Stojković, and Marius Schmidt. High-resolution crystal structures of transient intermediates in the phytochrome photocycle. *Structure (London, England: 1993)*, 29(7):743–754.e4, July 2021. tex.ids= 2021carrillo.

[24] Doeke R. Hekstra. Emerging Time-Resolved X-Ray Diffraction Approaches for Protein Dynamics. *Annual Review of Biophysics*, 52(1):255–274, May 2023.

[25] Christopher M. Bishop and Hugh Bishop. *Deep Learning: Foundations and Concepts*. Springer International Publishing, Cham, 2024.

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity Mappings in Deep Residual Networks, July 2016.

[27] Doeke R. Hekstra, Harrison K. Wang, Margaret A. Klureza, Jack B. Greisman, and Kevin M. Dalton. Sensitive Detection of Structural Differences using a Statistical Framework for Comparative Crystallography, July 2024. Pages: 2024.07.22.604476 Section: New Results.

[28] J. B. Greisman, K. M. Dalton, C. J. Sheehan, M. A. Klureza, I. Kurinov, and D. R. Hekstra. Native SAD phasing at room temperature. *Acta Crystallographica Section D: Structural Biology*, 78(8):986–996, August 2022. Publisher: International Union of Crystallography.

[29] Graeme Winter, James Beilsten-Edmands, Nicholas Devenish, Markus Gerstel, Richard J. Gildea, David McDonagh, Elena Pascal, David G. Waterman, Benjamin H. Williams, and Gwyndaf Evans. DIALS as a toolkit. *Protein Science*, 31(1):232–250, January 2022.

[30] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

A Plots

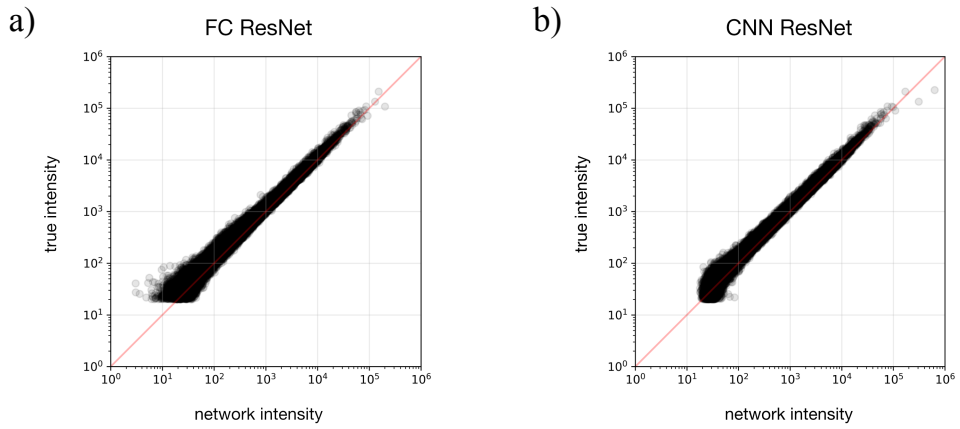

Figure 5: Correlation plots of the ground truth (simulated intensity) and network predictions (network intensity). Points that lie on the equality line are very close approximations to the ground truth.

Table 1: Training hyperparameters for experimental data

Hyperparameter	**Value**
channels (C)	6
height (H)	21
width (W)	21
number of images per shoebox (ϕ)	3
batch size	128
number of reflections	1,900,000
epochs	10
learning rate	0.001
p_I	Exponential
p_I rate	1.0
κ_1	0.0001
p_{Bg}	Exponential
p_{Bg} rate	1.0
κ_2	0.0001
q_I	Gamma
q_{Bg}	Gamma

Table 2: Comparison of R-work and R-free values across different models

Model	R-work		R-free	
	Start	Final	Start	Final
FC/MVN	0.2469	0.2175	0.2518	0.2424
CNN/MVN	0.2252	0.1912	0.2337	0.2170
CNN/Softmax	0.2234	0.1851	0.2299	0.2109
CNN/DIALS	0.2049	0.1628	0.2036	0.1776
DIALS	0.1982	**0.1482**	0.1889	**0.1530**

Table 3: RMSD values for anomalous peaks across different models

Peak Coordinates	CNN/DIALS	CNN/MVN	FC/MVN	DIALS	CNN/Softmax
(31.29, 8.123, 2.067)	28.22	9.28	-	**33.82**	10.89
(27.68, 27.68, 18.9)	26.80	6.95	8.79	**32.41**	10.70
(39.11, 9.928, 14.76)	14.79	-	-	**17.19**	5.66
(27.98, 10.53, 17.12)	13.56	-	-	**17.05**	-