Self-Supervised Deep Learning for Model Correction in the Computational Crystallography Toolbox

Vidya Ganapati [1 2] **Daniel Tchoń** [1] **Aaron S. Brewster** [1] **Nicholas K. Sauter** [1]

Abstract

The Computational Crystallography Toolbox (CCTBX) is open-source software that allows for processing of crystallographic data, including from serial femtosecond crystallography (SFX), for macromolecular structure determination. We aim to use the modules in CCTBX to determine the oxidation state of individual metal atoms in a macromolecule. Changes in oxidation state are reflected in small shifts of the atom's X-ray absorption edge. These energy shifts can be extracted from the diffraction images recorded in serial femtosecond crystallography, given knowledge of a forward physics model. However, as the diffraction changes only slightly due to the absorption edge shift, inaccuracies in the forward physics model make it extremely challenging to observe the oxidation state. We describe the potential impact of using self-supervised deep learning to correct the scientific model in CCTBX and provide uncertainty quantification. We provide code for forward model simulation and data analysis, built from CCTBX modules, at https://github.com/gigantocypris/SPREAD. Open questions in algorithm development are described to help spur advances through dialog between crystallographers and machine learning researchers. New methods could help elucidate charge transfer processes in many reactions, including key events in photosynthesis.

1. Introduction

Crystallography is a branch of science that revolves around investigating the internal structure of crystals. A typical crystal is comprised of a series of almost-identical unit cells, repeated periodically in all directions. Mathematically, it can be described as a convolution of a single average unit cell with a periodic three-dimensional lattice. The distribution of electron density in any crystal can be understood as a three-dimensional periodic wave in direct (experimental) space.

X-ray diffraction (XRD) is a common crystallographic technique used to determine the structure of crystals. In an XRD experiment, incident radiation is scattered by the electron density and is subsequently imaged on a detector. The collected diffraction pattern is related to the Fourier transform of the scatterer density. By the convolution theorem, the Fourier transform is a product of the Fourier transforms of the periodic crystal lattice and the electron density in a single unit cell. In the reciprocal space, this results in a set of discrete peaks that represent the crystal structure. Every peak is indexed using Miller indices $\vec{h} = [h\ k\ l]^\top$ and carries some information about the crystal structure in the form of a structure factor $F_{\vec{h}}$. Due to the imperfection of real crystals, the peaks in the reciprocal space are not infinitesimally small as expected from theory but rather slightly diffused. This effect can be modeled by treating the crystal as a finite set of small mosaic domains, each slightly misaligned relative to others.

For any fixed orientation, a detector images a two-dimensional spherical slice of the reciprocal space called the Ewald sphere. Rotating the crystal in the direct space rotates its Fourier transform in the reciprocal space, allowing the Ewald sphere to pass through different reciprocal space peaks, and deposit the information about their shape and intensity in the form of diffraction spots. The intensity of each diffraction spot, summed incoherently across all mosaic domains, is proportional to the modulus squared of its structure factor $F_{\vec{h}}$. The core problem of XRD structure determination is to retrieve structure factors $F_{\vec{h}}$ based on the intensities of all diffraction spots observed on a detector (Giacovazzo, 2011).

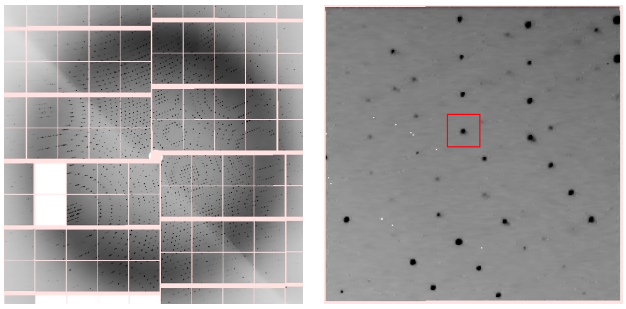
The Computational Crystallography Toolbox (CCTBX), is open-source software used to process data collected in crystallographic experiments to determine $|F_{\vec{h}}|^2$ of all the diffraction spots in the Fourier transform (Grosse-Kunstleve et al., 2002). Documentation and code are available at https://cci.lbl.gov/docs/cctbx and https://github.com/cctbx/cctbx_project, respectively. The resulting $|F_{\vec{h}}|^2$ can be processed by another tool such as PHENIX (Liebschner et al., 2019) in order to solve for the electron density of the macromolecule. The package CCTBX models the formation of

diffraction images using knowledge of the underlying imaging physics. From the collected data, the inverse problem of finding the underlying structure factor amplitudes can be solved.

Building off of CCTBX, we can model how the electronic state of metal atoms in a macromolecule informs the resulting crystal diffraction image (Sauter et al., 2020). Our scientific aim is to understand charge transfers at the atom level in photosystem II, a key protein complex in photosynthesis (Bhowmick et al., 2023). Such knowledge can inform the future development of solar fuels. We make our code available at `https://github.com/gigantocypris/SPREAD`. Here, we describe model limitations and the potential of machine learning to help answer scientific questions.

2. Serial Femtosecond Crystallography (SFX)

In a classical diffraction experiment, a single crystal is affixed on a goniometer and cooled down to a cryogenic temperature to limit X-ray radiation damage. Exposure over a series of orientations allows for collection of a complete set of spot intensities. However, the information collected in cryogenic conditions describes a structure far from its natural state.



(a) (b)

Figure 1. (a) Center portion of still shot of photosystem II in an SFX experiment. (b) Example shoebox drawn by DIALS.

Serial femtosecond crystallography (SFX) with X-ray free electron laser (XFEL) pulses allows for room-temperature measurements. In this modality, many micro-crystals are imaged sequentially using short X-ray pulses. A single pulse is used to capture a single "still shot", recording a diffraction image of a crystal right before it is destroyed by the radiation. The measurements can be synchronized with an auxiliary laser, allowing observation of short-lived transient states (pump-probe experiments) (Kern et al., 2012; 2013). All still shots are assumed to originate from crystals with identical chemical composition, but their orientation, exact unit cell lengths, and mosaicity parameters may vary from sample to sample (Mendez et al., 2020).

Each still shot is composed of pixels collected from multiple detector panels, and the detector is described in a hierarchical manner (Brewster et al., 2018), see Fig. 1a. The Diffraction Integration for Advanced Light Sources (DIALS) software package (Winter et al., 2018) is built on top of CCTBX and can perform diffraction spot finding, see Fig. 1b. The documentation is available at `https://dials.github.io` and the code at `https://github.com/dials/dials`. With SFX, the orientation of the crystal in a still shot, as well as an overall scaling factor that depends on variations in the incident beam and volume of illuminated crystal, are initially unknown (Evans & Murshudov, 2013). The unit cell parameters also vary from crystal to crystal. After spot finding, DIALS determines the orientation of the crystal and assigns a Miller index to each spot. The structure factor amplitudes are then found by integrating, scaling, and merging the spot intensity values over all still shots (Winter et al., 2018).

3. Refinement with NANOBRAGG

No crystal is a perfectly periodic monolith; rather, it can be better described as a cluster of small periodic domains. Individual crystalline domains are very similar, but can slightly vary in their orientation and shape. Every diffraction spot originating from the crystal is a sum of the intensity contributions of each mosaic domain. Likewise, the resulting diffraction spots from a polychromatic spectrum can be modeled as the superposition of spots from each wavelength. Both the incident spectrum and mosaicity are not considered in DIALS, as the software integrates every spot without consideration of its shape. Diffraction spot shape can, however, be modeled with NANOBRAGG, a module in CCTBX (Holton et al., 2014; Lyubimov et al., 2016).

During integration, DIALS investigates a small region of the still shot around each spot called a "shoebox." To solve the inverse problem of determining the structure factors as well as other parameters such as orientation and unit cell, the shoeboxes can be simulated from the underlying parameters with NANOBRAGG. The simulation can be then compared with the experimental data to determine a loss function. The structure factors and shoebox-dependent parameters can be updated by the gradients with respect to a loss function until convergence (Mendez et al., 2020). The structure factor and parameter estimates from DIALS are used as initial conditions to make the inverse problem computationally tractable. We note that with both DIALS and NANOBRAGG, structure factor amplitudes are point estimates; there is no uncertainty quantification.

4. Spatially Resolved Anomalous Dispersion (SPREAD)

Building off of NANOBRAGG and CCTBX, we can determine the oxidation state of individual metal atoms in a macromolecule with data from SFX. Changes in oxidation state are reflected in slight shifts (on the order of 1-2 electron volts) of the atom's *K* absorption edge. These shifts are embedded in SFX data. There is a wavelength dependency in the structure factor $F_{\vec{h}}$ as the scattering factor of each atom in the macromolecule includes a complex wavelength-dependent quantity known as the anomalous scattering factor (Sauter et al., 2020). Far from the *K* absorption edge, the scattering factor is approximately constant over wavelength. We aim to solve for the anomalous scattering factor for atoms with K edge near the center wavelength of the incident spectrum. This technique, known as spatially resolved anomalous dispersion (SPREAD), can yield insight into electron movement during a chemical reaction. In particular, our scientific aim is to solve for the anomalous scattering factors of the four manganese (Mn) atoms in photosystem II to elucidate single-electron transfers in photosynthesis (Sauter et al., 2020).

The structure factor at the Miller index $\vec{h}$ is given as the sum of contributions from each atom m of the macromolecule:

$$F_{\vec{h}}(\lambda) = \sum_m F_{\vec{h},m}(\lambda), \tag{1}$$

where λ denotes the wavelength of incident radiation. As described in (Sauter et al., 2020), the wavelength dependent contribution from each atom, $F_{\vec{h},m}$, can be expressed as:

$$F_{\vec{h},m}(\lambda) = \left[f_m^0(|\vec{Q}|) + \Delta f_m'(\lambda) + i\Delta f_m''(\lambda)\right] \times \exp\left[2\pi i \vec{r_m} \cdot \vec{h}\right] \times \exp(-B_m|\vec{Q}|^2/4), \tag{2}$$

where $\Delta f_m'$ and $\Delta f_m''$ are the real and imaginary parts of the wavelength-dependent anomalous scattering factor for the m^{th} atom, related by the Kramer's Kronig relationship (Meurer et al., 2022; Sherrell, 2014). The anomalous scattering factor changes with valence state, but is constant over the magnitude of the scattering vector $|\vec{Q}|$. The non-anomalous scattering factor of the atom is denoted by f_m^0 and depends on the scattering vector $|\vec{Q}|$. The position of the atom within the unit cell using fractional coordinates is $\vec{r_m}$. The Miller index is denoted by $\vec{h}$, while B_m is the atom's temperature-dependent B factor. For an incident spectrum centered at 6550 eV, the $\Delta f_m'$ and $\Delta f_m''$ terms are negligible for all photosystem II atoms, except the four Mn atoms. The functions $\Delta f_m'(\lambda)$ and $\Delta f_m''(\lambda)$ shift by a few electron volts between manganese in its $3+$ and $4+$ oxidation state; this is the change we aim to determine.

The change in the total structure factor due to a change in the electronic state of a few constituent atoms is small. There are thus strict requirements on the accuracy of the forward physics model. For example, an inaccurate description of crystal mosaicity may lead to an incorrect determination of the anomalous scattering factors. So far, SPREAD with SFX data has only been performed successfully with simulated data, with code that extends NANOBRAGG and CCTBX (Sauter et al., 2020). This prior work on simulated data results in point estimates of the anomalous scattering factor as a function of wavelength for the atoms of interest. Application of the methods to real SFX data has not been successful thus far. We describe the potential impact of using self-supervised deep learning to correct the scientific model and provide uncertainty quantification.

5. Model Correction with Neural Networks

In an inverse problem, we have measurements and a known forward physics model. Our aim is to discover the source of those measurements. If we know the source and the forward physics model (the forward problem), determining the probability distribution of measurements is straightforward. However, the inverse problem of determining the source given the measurements is more challenging and may be ill-posed. If we have a fully specified forward model, we can take an optimization-based approach, using gradient descent to tune an initial guess of the source in order to maximize the likelihood of achieving the given measurements. The inverse problem becomes even more challenging if we have incomplete or incorrect knowledge of the forward model. The complete, correct form of the forward model may be unknown due to experimental unknowns and complexities. If we have measurements on multiple sources, with each measurement obeying the same underlying (incomplete) forward model, we may have enough information to both correct the model and determine all the sources. In this case, our optimization objective is to maximize the total likelihood of all measurements, with a penalty for deviating too far from the known forward physics. Here, we describe related work, outline our framework for model correction in SFX, and discuss open questions.

5.1. Related Work

5.1.1. CARELESS

The software package CARELESS (Dalton et al., 2022) provides a model-free way to correct structure factor amplitudes $|F_{\vec{h}}|$ derived from DIALS. Each structure factor from DIALS is assumed to be the product of an image-dependent, spot-dependent scale factor, and the true structure factor amplitude. The prior distribution on the true structure factor amplitude is given; the prior on the scale factor is assumed to be uninformative. A neural network takes metadata on

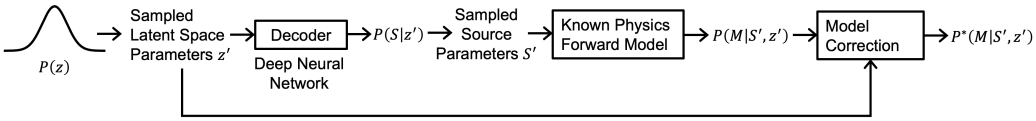


Figure 2. Generator for source parameters; model correction could be performed with a normalizing flow (Kobyzev et al., 2021).

the diffraction spot (e.g. crystal orientation, location on the detector, image number, and Miller index), and outputs a prediction of the posterior probability distribution of the corresponding scale factor. Variational inference is used to train this neural network as well as the parameters of the distribution estimating the structure factor amplitude posterior, pushing both distributions to the true posteriors. Intuitively, variational inference attempts a balance between maximizing the likelihood of the measured intensity data while not veering far from the prior distribution on the structure factor $|F_{\vec{h}}|$. The metadata used as input to the neural network must be chosen judiciously, as it is possible for the scale factor to overexplain the experimental $|F_{\vec{h}}|$, creating a poor true structure factor amplitude estimate.

Like CARELESS, we aim to perform model correction. However, our goal is primarily to find anomalous scattering factors of certain atoms; for SPREAD, we assume the structure factor amplitude is known. For this task, our framework needs to take into account pixel-by-pixel variations in diffraction spot shoeboxes, as well as utilize the (partially) known underlying physics.

5.1.2. PHYSICS-INFORMED VARIATIONAL AUTOENCODER (P-VAE)

Recent work has evaluated the use of physics-informed variational autoencoders (P-VAEs) to solve inverse problems in imaging (Mendoza et al., 2022; Olsen et al., 2022); a similar formulation is described in (Leong et al., 2023). Specifically, these works consider a dataset of measurements, with each measurement on a different unknown source. Due to experimental limitations, the measurement on each source is sparse. There is not enough information in a single measurement M to recover the corresponding source S using conventional optimization methods for inverse problems. However, the entire dataset of measurements is large, i.e. there are sparse measurements on many similar sources. The P-VAE jointly solves for the underlying prior distribution on the sources $P(S)$ and all the posterior distributions on the dataset, $P(S|M)$. In SFX, we have a similar problem where we have sparse (i.e. single orientation) measurements on many similar but different crystals. The formulation of the P-VAE can assist in the determination of a probability distribution for the anomalous scattering factors, as opposed to yielding just a point estimate.

5.1.3. INCOMPLETE FORWARD MODELS

The frameworks in (Mendoza et al., 2022; Olsen et al., 2022; Leong et al., 2023) focus on the problem of sparse measurements, they do not consider an incomplete forward model. A partially specified forward physics model in the context of a P-VAE is considered in (Takeishi & Kalousis, 2021). A neural network is trained to transform the incomplete model into the completed one. The augmented forward model is penalized for veering from the known incomplete forward model through additional loss terms added to the P-VAE loss. To solve for the anomalous scattering factors in SFX, we face the dual problems of sparse measurements and an incomplete forward model.

5.2. P-VAE for SPREAD

Here, we outline how P-VAEs could be applied to processing SFX data for SPREAD and describe open questions. Code and documentation for the SPREAD forward model are given at `https://github.com/gigantocypris/SPREAD`.

We process collected still shots with DIALS, drawing shoeboxes around diffraction spots and obtaining estimates for unit cell shape, orientation, and overall scale factor. We aim to find a probability distribution of the underlying anomalous scattering factor functions $\Delta f'_m(\lambda)$ and $\Delta f''_m(\lambda)$. To do so, we can create a "generator" with a latent space z that can be sampled to yield shoebox source parameters from the same underlying distribution as the measured shoeboxes. The generator includes the partially known forward physics model and model correction; see Fig. 2.

We want to train the generator to maximize the probability of obtaining the actual shoeboxes. This problem is made computationally tractable by using an encoder that takes an actual shoebox and its metadata as input, and approximates the conditional probability distribution $P(z|\text{shoebox})$. Connecting the encoder to the generator creates a P-VAE. The derivation of the P-VAE loss function with model correction is given in Appendix A. Training the networks with the P-VAE loss recovers the distribution governing the anomalous scattering factors $\Delta f'_m(\lambda)$ and $\Delta f''_m(\lambda)$.

The data of a single shoebox can be represented in a hierarchical manner, with latent parameters shared amongst all shoeboxes at the dataset and image levels. We outline the basic framework of a hierarchical P-VAE in Fig. 3, with

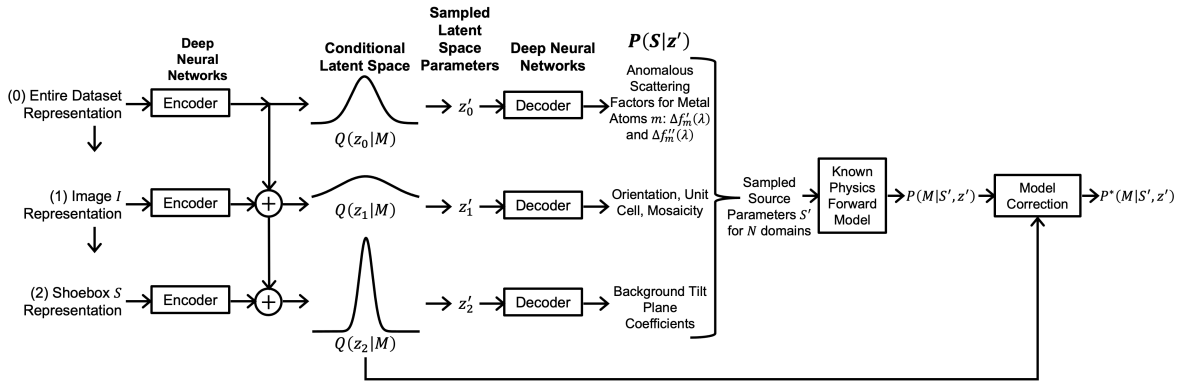


Figure 3. Hierarchical P-VAE for determining anomalous scattering factors.

further details in Appendix A. We describe how a similar framework can be used to refine structure factor amplitudes in Appendix B.

5.3. Open Questions

We describe a potential framework to use machine learning for model correction in SFX. However, there are many open questions, such as:

- What trade-off do we make between following the scientific model generated by first principles and allowing model corrections with deep neural networks? Relatedly, how do we verify correctness?
- Conventional crystallographic data analysis rejects a significant portion of collected data. Can deep learning techniques allow for insight to be extracted from poor-quality data?
- What neural network architecture is needed for model correction? How sensitive is the procedure to neural network architecture?
- How do we best incorporate the results from DIALS and NANOBRAGG for quantities such as unit cell, orientation, scale, and mosaicity?

6. Conclusion

We describe the Computational Crystallography Toolbox (CCTBX) and the potential for applying self-supervised physics-informed deep learning methods for analysis in serial femtosecond crystallography (SFX). A scientific problem of interest in SFX is determining the anomalous scattering factors of specific metal atoms in macromolecules. Such knowledge will allow determination of the oxidation state of individual metal atoms in the macromolecule, elucidating charge transfer processes in chemical reactions.

We outline the potential use of deep neural networks to make arbitrary corrections to the forward model in SFX, supplementing previous work (Sauter et al., 2020; Mendez et al., 2020; Brehm et al., 2023) that solely optimizes variables that parametrize a forward model derived from first principles. The goal is to correct for experimental effects of unknown origin, relaxing the stringent model accuracy requirements for spatially resolved anomalous dispersion (SPREAD). These methods have the potential to improve data analysis, with the impact of discovering new science by striking a balance between knowledge of an ideal forward model, and knowledge learned from data. We present open questions to facilitate collaborations between crystallographers and deep learning researchers, with the aim of accelerating progress in SFX. Code for forward model simulation with CCTBX modules and instructions for conventional analysis with DIALS are given at `https://github.com/gigantocypris/SPREAD`.

Acknowledgements

The authors acknowledge Daniel W. Paley and Iris D. Young for useful discussions and the still shot of photosystem II. N.K.S. acknowledges support from the National Institutes of Health (NIH) grant R01-GM117126, and the Exascale Computing Project (grant 17-SC20-SC), a collaborative effort of the DOE Office of Science and the National Nuclear Security Administration.

References

Bhowmick, A., Hussein, R., Bogacz, I., Simon, P. S., Ibrahim, M., Chatterjee, R., Doyle, M. D., Cheah, M. H., Fransson, T., Chernev, P., Kim, I.-S., Makita, H., Dasgupta, M., Kaminsky, C. J., Zhang, M., Gätcke, J., Haupt, S., Nangca, I. I., Keable, S. M., Aydin, A. O., Tono, K., Owada, S., Gee, L. B., Fuller, F. D., Batyuk, A., Alonso-Mori, R., Holton, J. M., Paley, D. W., Moriarty, N. W., Mamedov, F., Adams, P. D., Brewster, A. S., Dobbek,

H., Sauter, N. K., Bergmann, U., Zouni, A., Messinger, J., Kern, J., Yano, J., and Yachandra, V. K. Structural evidence for intermediates during O2 formation in photosystem II. *Nature*, 617(7961):629–636, May 2023.

Brehm, W., White, T., and Chapman, H. N. Crystal diffraction prediction and partiality estimation using Gaussian basis functions. *Acta Crystallographica Section A Foundations and Advances*, 79(2):145–162, 2023.

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

Dalton, K. M., Greisman, J. B., and Hekstra, D. R. A unifying Bayesian framework for merging X-ray diffraction data. *Nature Communications*, 13(1):7764, 2022.

Doersch, C. Tutorial on Variational Autoencoders. *arXiv:1606.05908 [cs, stat]*, 2016.

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

Giacovazzo, C. (ed.). *Fundamentals of crystallography*. Oxford University Press, Oxford ; New York, 3rd ed edition, 2011.

Grosse-Kunstleve, R. W., Sauter, N. K., Moriarty, N. W., and Adams, P. D. The *Computational Crystallography Toolbox* : crystallographic algorithms in a reusable software framework. *Journal of Applied Crystallography*, 35 (1):126–136, 2002.

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

Kern, J., Alonso-Mori, R., Hellmich, J., Tran, R., Hattne, J., Laksmono, H., Glöckner, C., Echols, N., Sierra, R. G., Sellberg, J., Lassalle-Kaiser, B., Gildea, R. J., Glatzel, P., Grosse-Kunstleve, R. W., Latimer, M. J., McQueen, T. A., DiFiore, D., Fry, A. R., Messerschmidt, M., Miahnahri, A., Schafer, D. W., Seibert, M. M., Sokaras, D., Weng, T.-C., Zwart, P. H., White, W. E., Adams, P. D., Bogan, M. J., Boutet, S., Williams, G. J., Messinger, J., Sauter, N. K., Zouni, A., Bergmann, U., Yano, J., and Yachandra, V. K. Room temperature femtosecond X-ray diffraction of photosystem II microcrystals. *Proceedings of the National Academy of Sciences of the United States of America*, 109(25):9721–9726, 2012.

Kern, J., Alonso-Mori, R., Tran, R., Hattne, J., Gildea, R. J., Echols, N., Glöckner, C., Hellmich, J., Laksmono, H., Sierra, R. G., Lassalle-Kaiser, B., Koroidov, S., Lampe, A., Han, G., Gul, S., Difiore, D., Milathianaki, D., Fry, A. R., Miahnahri, A., Schafer, D. W., Messerschmidt, M., Seibert, M. M., Koglin, J. E., Sokaras, D., Weng, T.-C., Sellberg, J., Latimer, M. J., Grosse-Kunstleve, R. W., Zwart, P. H., White, W. E., Glatzel, P., Adams, P. D., Bogan, M. J., Williams, G. J., Boutet, S., Messinger, J., Zouni, A., Sauter, N. K., Yachandra, V. K., Bergmann, U., and Yano, J. Simultaneous femtosecond X-ray spectroscopy and diffraction of photosystem II at room temperature. *Science*, 340(6131):491–495, 2013.

Kingma, D. P. and Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]*, 2014.

Kobyzev, I., Prince, S. J. D., and Brubaker, M. A. Normalizing Flows: An Introduction and Review of Current Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):3964–3979, 2021. arXiv:1908.09257 [cs, stat].

Leong, O., Gao, A. F., Sun, H., and Bouman, K. L. Ill-Posed Image Reconstruction Without an Image Prior, 2023. arXiv:2304.05589 [cs, eess].

Liebschner, D., Afonine, P. V., Baker, M. L., Bunkóczi, G., Chen, V. B., Croll, T. I., Hintze, B., Hung, L.-W., Jain, S., McCoy, A. J., Moriarty, N. W., Oeffner, R. D., Poon, B. K., Prisant, M. G., Read, R. J., Richardson, J. S., Richardson, D. C., Sammito, M. D., Sobolev, O. V., Stockwell, D. H., Terwilliger, T. C., Urzhumtsev, A. G., Videau, L. L., Williams, C. J., and Adams, P. D. Macromolecular structure determination using X-rays, neutrons and electrons: recent developments in *Phenix*. *Acta Crystallographica Section D Structural Biology*, 75(10):861–877, 2019.

Lyubimov, A. Y., Uervirojnangkoorn, M., Zeldin, O. B., Zhou, Q., Zhao, M., Brewster, A. S., Michels-Clark, T., Holton, J. M., Sauter, N. K., Weis, W. I., and Brunger, A. T. Advances in X-ray free electron laser (XFEL) diffraction data processing applied to the crystal structure of the synaptotagmin-1 / SNARE complex. *eLife*, 5: e18740, 2016.

Mendez, D., Bolotovsky, R., Bhowmick, A., Brewster, A. S., Kern, J., Yano, J., Holton, J. M., and Sauter, N. K. Beyond integration: modeling every pixel to obtain better structure factors from stills. *IUCrJ*, 7(6):1151–1167, 2020.

Mendoza, R., Nguyen, M., Zhu, J. W., Dumont, V., Perciano, T., Mueller, J., and Ganapati, V. A Self-Supervised Approach to Reconstruction in Sparse X-Ray Computed Tomography, 2022. arXiv:2211.00002 [physics].

Meurer, F., Dolomanov, O. V., Hennig, C., Peyerimhoff, N., Kleemiss, F., Puschmann, H., and Bodensteiner, M. Refinement of anomalous dispersion correction parameters in single-crystal structure determinations. *IUCrJ*, 9(5): 604–609, 2022.

Olsen, A., Hu, Y., and Ganapati, V. Data-Driven Computational Imaging for Scientific Discovery, 2022. arXiv:2210.16709 [physics].

Sauter, N. K., Kern, J., Yano, J., and Holton, J. M. Towards the spatial resolution of metalloprotein charge states by detailed modeling of XFEL crystallographic diffraction. *Acta Crystallographica Section D Structural Biology*, 76 (2):176–192, 2020.

Sherrell, D. Diffraction spectroscopy of metalloproteins. 2014. PhD Thesis, University of Saskatchewan.

Takeishi, N. and Kalousis, A. Physics-Integrated Variational Autoencoders for Robust and Interpretable Generative Modeling, 2021. arXiv:2102.13156 [cs, stat].

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

A. Derivation of Physics-Informed Variational Autoencoder (P-VAE) Loss

In a variational autoencoder (Kingma & Welling, 2014; Doersch, 2016), the goal is to learn how to generate new examples, sampled from the same underlying probability distribution as a training dataset of m sources $\{S_1, S_2, ...S_m\}$. In our case, a single source S fully specifies a diffraction spot shoebox with underlying parameters including orientation, unit cell, mosaicity, anomalous scattering factors of metal atoms, parameters characterizing the background, and model correction terms. To accomplish the task of creating a shoebox generator, a latent random variable z is created that describes the space on a lower-dimensional manifold. A deep neural network defines a function (the "decoder") from a sample of z to a conditional probability distribution $P(S|z)$, see Fig. 4.

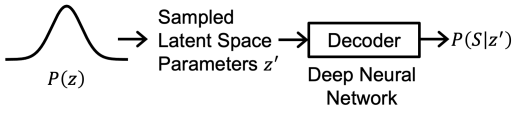


Figure 4. Generator for sources S.

The parameters of the decoder network are optimized to maximize the probability of generating the independent sources of the training dataset:

$$\log P(S_1, S_2, ...S_m) = \sum_{i=1}^{m} \log P(S_i) = \sum_{i=1}^{m} \log \left[\int P(S_i|z)P(z)dz \right] \tag{3}$$

However, we do not have any ground truth sources S, but rather a dataset of noisy diffraction spot shoebox measurements $\{M_1, M_2, ...M_m\}$. Each measurement M consists of the pixel intensity values inside the shoebox. If we assume the forward model $P(M|S)$ is known, instead of maximizing the probability of generating S, we can maximize the probability of generating the measurements:

$$\log P(M_1, M_2, ...M_m) = \sum_{i=1}^{m} \log P(M_i) = \sum_{i=1}^{m} \log \left[\int \int P(M_i|S)P(S|z)P(z)dSdz \right] \tag{4}$$

This modified "physics-informed" generator is seen in Fig. 5.

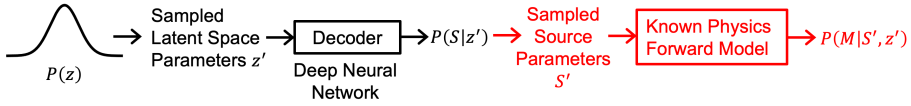


Figure 5. Physics-informed generator for sources S. The modifications to a conventional generator are highlighted in red.

If the forward model is only partially specified, the likelihood $P(M|S)$ can be modified to $P^*(M|S)$ by processing through a normalizing flow (Kobyzev et al., 2021), see Fig. 2.

A penalty for deviating from the known forward model can be added to the overall optimization objective, such as a term proportional to the Kullback–Leibler (KL) divergence between the modified and unmodified likelihoods, to force the generator to try to first explain the data with physics before applying a trainable modification.

We note that different independent latent variables can separately underlie quantities such as unit cell, mosaic shape and size, and orientation. The latent variables can be organized hierarchically: global over the entire dataset (e.g. anomalous scattering factors), per image (e.g. orientation, incident photon spectrum), and per shoebox (e.g. background noise parameters). The crystal is composed of many mosaic domains, each with a different unit cell, shape, size, and mis-orientation; the set of source parameters can be sampled N times, where N is the number of total mosaic domains modeled. The hierarchical generator can output shoeboxes with knowledge of the pixel positions on the detector, see Fig. 6.

Considering a single example (as in stochastic training with batch size of 1), we aim to maximize $\log P(M)$; for most sampled values of z' and S', the probability $P(M|S', z')$ is close to zero, causing poor scaling of sampled estimates to

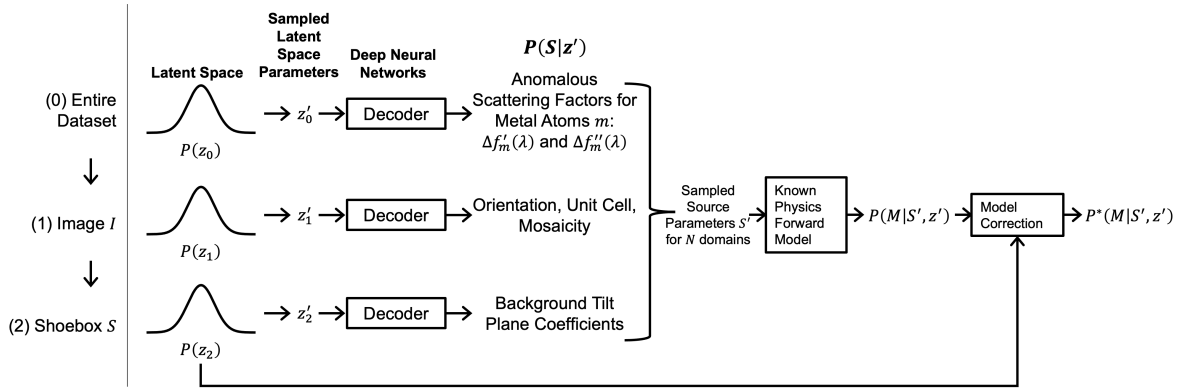


Figure 6. Hierarchical physics-informed generator with model correction for sources S.

the integral $\int\int P(M|S)P(S|z)P(z)dSdz$. We follow the P-VAE formulation (Mendoza et al., 2022; Olsen et al., 2022), estimating the parameters of $P(z|M)$ by processing the measurement M using a deep neural network called the "encoder," see Fig. 7.

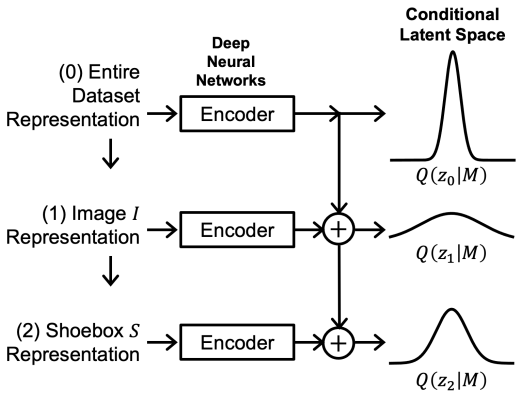


Figure 7. Hierarchical encoder for a P-VAE.

The output of the encoder is an estimate of $P(z|M)$ denoted as $Q(z|M)$. The KL divergence between the distributions is given by:

$$D_{KL}[Q(z|M)||P(z|M)] = E_{z\sim Q}[\log Q(z|M) - \log P(z|M)] \tag{5}$$

We also have, by Bayes' Theorem:

$$\log P(z|M) = \log P(M|z) + \log P(z) - \log P(M) \tag{6}$$

Combining the expressions yields:

$$\log P(M) - D_{KL}[Q(z|M)||P(z|M)] = E_{z\sim Q}\left[\log \int P(M|S)P(S|z)dS\right] - D\left[Q(z|M)||P(z)\right]. \tag{7}$$

The first term on the right side of this expression can be estimated with a sample-based estimate. As KL divergence is always ≥ 0 and reaches 0 when $Q(z|M) = P(z|M)$, maximizing the right side during training causes $P(M)$ to be maximized while forcing $Q(z|M)$ towards $P(z|M)$. When forward model correction is applied, a term can be added to penalize the distance between $P(M|S)$ and $P^*(M|S)$. The full physics-informed variational autoencoder is visualized in Fig. 3.

B. Framework for Structure Factor Refinement

Our focus in this paper is the determination of anomalous scattering factors. Machine learning has great potential for impact in this area, as there are stringent forward model accuracy requirements. In the solution of anomalous scattering factors, we assume knowledge of the structure of the macromolecule. However, refinement of the structure factor amplitudes is also possible from a similar framework, see Fig. 8.

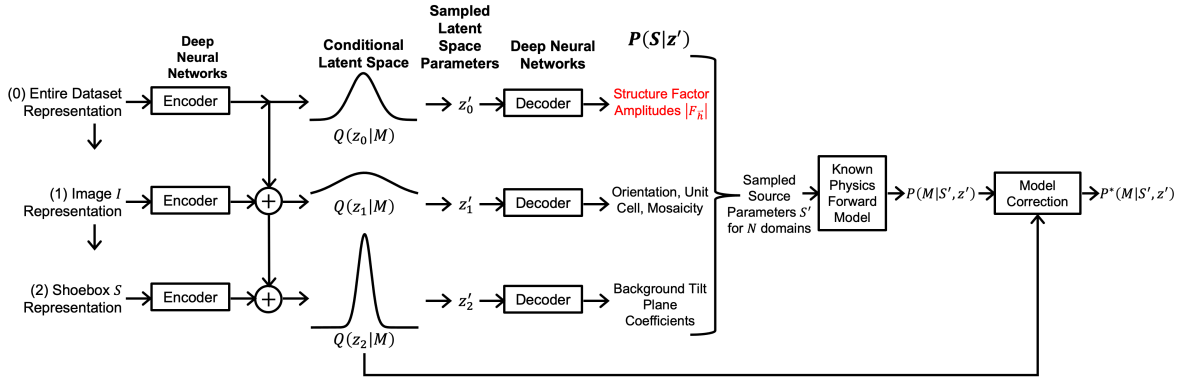


Figure 8. Hierarchical P-VAE for determining structure factor amplitudes.

Initial estimates of structure factor amplitude as well as orientation and unit cell can be found with DIALS and NANOBRAGG. These values can be used for initial training of the encoder and decoder, allowing a warm-start to the training procedure.