Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Learning the Crystallographic Phase Problem

Project Report
Project Dates: 27/07/2021 - 10/09/2021
University of Edinburgh

Project Student: Sarah Jane Scripps, University of Edinburgh
Supervised by: James Cumby, University of Edinburgh

Report Date: 18/10/2021

AI3SD-Intern-Series:Report-8_Scripps

Learning the Crystallographic Phase Problem
AI3SD-Intern-Series:Report-8_Scripps
Report Date: 18/10/2021
DOI: 10.5258/SOTON/AI3SD0149
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Project Details

Title	Learning the Crystallographic Phase Problem
Project Reference	AI3SD-FundingCall3_009
Supervisor Institution	University of Edinburgh
Project Dates	27/07/2021 - 10/09/2021
Website	Functional Materials Group Website
Keywords	Crystallography, phase problem, neural network, diffraction

2 Project Team

2.1 Project Student

Name and Title	Miss Sarah Jane Scripps
Employer name / University Department Name	University of Edinburgh, School of Chemistry
Work Email	s1904252@ed.ac.uk

2.2 Project Supervisor

Name and Title	James Cumby
Employer name / University Department Name	University of Edinburgh, School of Chemistry
Work Email	james.cumby@ed.ac.uk
Website Link	Functional Materials Group Website

2.3 Project Co-Supervisor

Name and Title	Dr Sohan Seth
Employer name / University Department Name	University of Edinburgh, School of Informatics
Work Email	sohan.seth@ed.ac.uk
Website Link	Personal website

3 Lay Summary

X-ray crystallography is the most common technique for the determination of the molecular structure in a crystal, essential to many areas including drug development, battery technology and determining the structure of viruses. Despite this importance, determining atomic structure using crystallography requires skilled experts to dedicate significant amounts of time and intuition.

X-rays interact with atoms via the phenomenon of diffraction, in which they are scattered by an atom's electrons. The scattered rays interfere and produce a pattern of light and dark spots, known as a diffraction pattern. During this process, some intrinsic information about the scattered rays is lost, making it mathematically impossible to explicitly calculate the structure of the crystal from the diffraction pattern. Current crystallographic methods centre around making an initial guess of the lost information, and then iteratively improving the result. The final structure is highly dependent on the initial guess, hence the need for experienced crystallographers to control the process. This project aimed to employ neural networks to predict crystal structures from diffraction patterns, with the aim of removing the need for human intervention and allowing faster, more accurate structure determination.

Two types of neural network were trialled with the problem. The networks were provided with sample diffraction patterns as input and the corresponding structures as target outputs. Both networks were then given a diffraction pattern not included in the training set and predicted a structure based on this input. Both were able to produce an output prediction. The better of the two networks was able to resolve atomic positions similar to the true structure, although with some distortion. This is a promising start for the technique, with further refinement of the network and training it is hopeful that greater accuracy can be achieved.

4 Aims and Objectives

X-ray diffraction patterns can be used to determine the unit cell dimensions and hence volume (V), reciprocal lattice coordinates (h, k, l) and the complex sum of the scattering factors (F_{hkl}). The symmetry of the unit cell can also be obtained.

The mathematical relationship between the electron density (ρ) of a periodic structure and the corresponding diffraction pattern is a Fourier transform, which produces a distribution of intensity in so-called "reciprocal" space. An inverse Fourier transform reproduces the distribution of electron density;

$$\rho(xyz) = \frac{2}{V} \sum_{hkl}^{\infty} |F_{hkl}| \cos 2\pi [hx + ky + lz - \phi_{hkl}]$$

In producing a diffraction pattern, all phase information (ϕ_{hkl}) of the scattered rays is lost and therefore cannot be measured. This means that the electron density map cannot be explicitly calculated from the information in a diffraction pattern, and is called the "phase problem" in crystallography. The phase problem is a limiting factor in current crystallographic methods. There are several existing approaches for solving this problem, typically relying on statistical refinement of the predicted structure. By altering initial phase assumptions and comparing to existing solutions for similar structures, the improvements should converge on the true structure. However, these methods are subject to phase bias. It is possible for an incorrect phase assumption or atomic assignment to be made which propagates through the subsequent refinement steps and can potentially lead to highly inaccurate structure predictions. The outputs of current methods are therefore not always accurate and cannot be relied upon without manual intervention.

The aim of this project was to determine whether neural networks could be used to predict the electron density maps from diffraction patterns, avoiding the traditional methods which can lead to phase bias. If successful, such a method could be employed across crystallographic disciplines to speed up the rate of structure solution.

5 Methodology

Input diffraction patterns and their corresponding target output structures were generated from existing structures in the Materials Project Database, via pre-existing functions developed by the Functional Materials group at the University of Edinburgh.

A modified diffraction pattern was used where diffracted intensity at each integer (h, k, l) was calculated using existing functions within the PyMatGen package, but with the atomic X-ray form factor replaced by the atomic number. 3D diffraction patterns (e.g. Fig. 1) were generated to a specified maximum value for h, k and l; in this project we used a maximum of either 10 (1000 points) or 20 (8000 points). Diffraction patterns were calculated for a subset of materials from the Materials Project, totalling ≈12,000 materials.

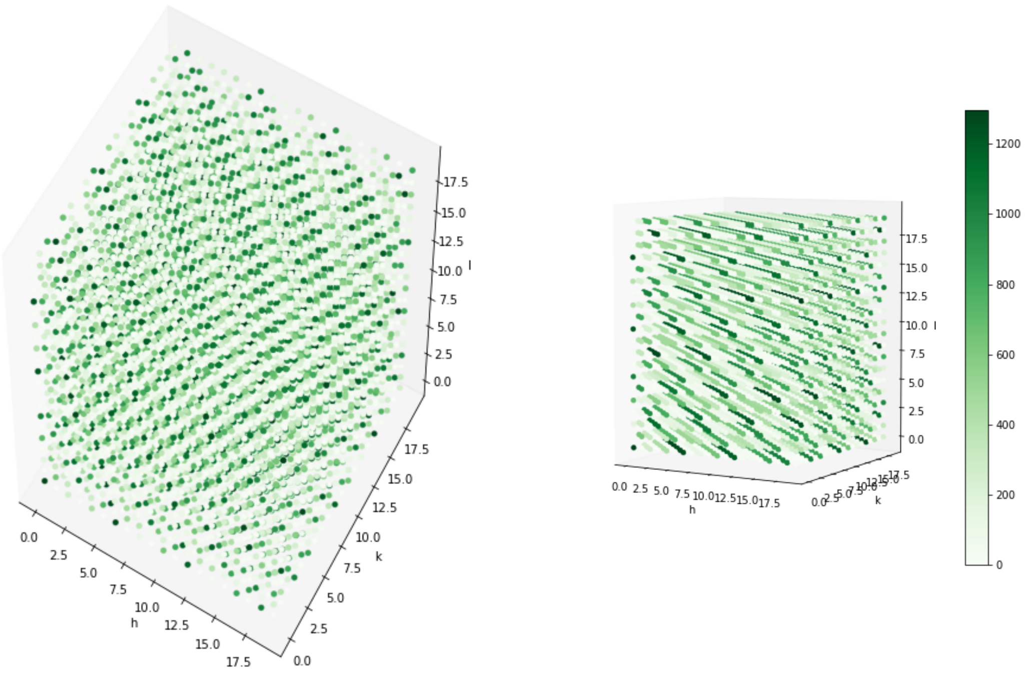


Figure 1: Two different orientations of the $20 \times 20 \times 20$ input diffraction pattern calculated for a fluorine crystal containing two F_2 molecules. Colour represents the diffracted intensity.

The point-like atomic structures (Fig. 2) were stored as voxelised distributions of the same dimension as the diffraction patterns, with a Gaussian spread applied to the atomic positions (Fig. 3). Voxelised structures were stored both in Cartesian coordinates (cuboidal voxels) and as fractions of the unit cell axes (voxels could be parallelepiped) to compare these two approaches.

The first network trialled was a convolutional neural network (CNN) — 3D-Unet -– as implemented here. The current usage of 3D-Unet is in medical imaging, [1] so the standard input data set is much larger than even the 20 x 20 x 20 diffraction pattern. Whilst diffraction intensities could be calculated to sufficiently large (h, k, l) indices to generate sufficient input, the inverse relationship between the magnitude of (h, k, l) and the interatomic distance would mean that the extra information would be largely meaningless, corresponding to distances shorter than the radius of a single atom. The existing 20 x 20 x 20 data were therefore up-sampled by duplication to meet the minimum patch size for 3D-Unet inputs.

The second network tested was a multi-layer perceptron (MLP) regressor, implemented within Scikit-Learn. Separate models were trained on 10 x 10 x 10 and 20 x 20 x 20 input dimensions; these are termed MLP10 and MLP20, respectively. In both cases, 100 hidden layers were utilised with ReLU activation. Otherwise, the default settings were used due to

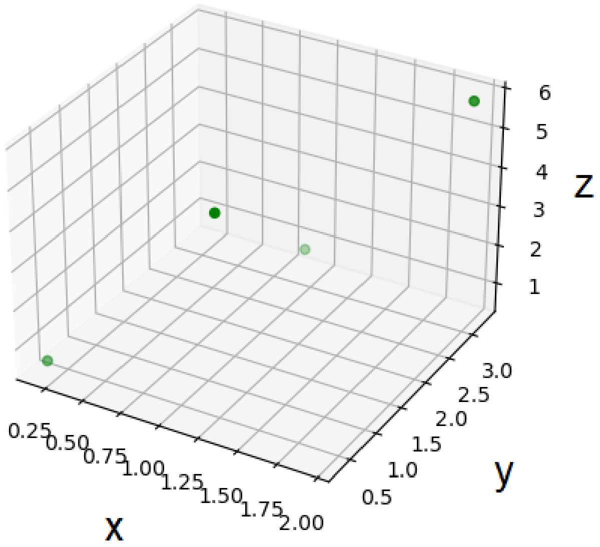


Figure 2: 3D visualisation of the atomic positions within the fluorine structure. Shading illustrates depth in the 3D Cartesian space.

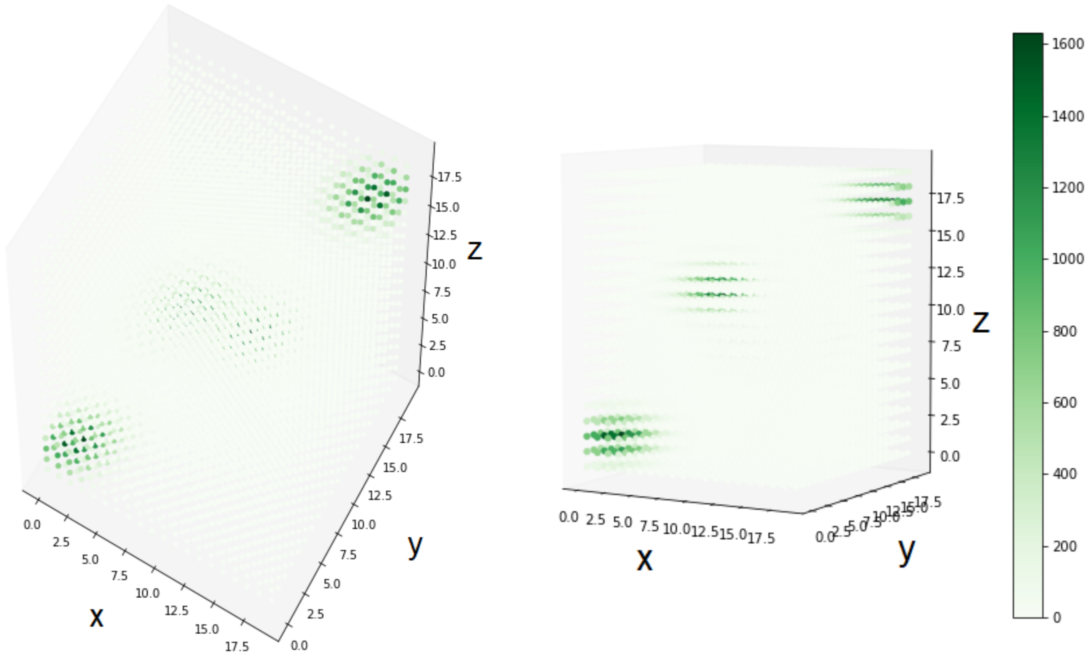


Figure 3: Two views of the voxelised structure for solid fluorine, showing the Gaussian spread applied to each atomic position w.r.t. Fig. 2.

time constraints.

For both models, training sets of varying sizes were used to train the network. Memory was a limiting factor in the case of 3D-Unet, leading to limited sample sizes of 56 and a maximum of 150 structures. The MLP Regressor was trained on a set of 150 structures initially, followed by increasing the training set to use the full database of 12,000 structures. The limited training sets were randomly selected from the full database. The smaller samples are a subset of the larger sets to ensure consistency.

Once training was completed the networks were used to predict a structure, which had not been included in the training set. For easiest comparison between configurations, sample sizes and networks the structure data used for predictions was kept consistent throughout. The output predictions are fractional voxelised distributions with the same output size as the input diffraction pattern. For direct comparison the 3D-Unet outputs are downsampled to a similar size as the MLP Regressor.

6 Results

6.1 3D-Unet

Unfortunately, the trained 3D-Unet model failed to resolve any obvious clusters of atomic density from the input diffraction data, although the predicted distributions do show a high degree of spatial symmetry. This can be seen in figure 4, where most voxels have some non-zero value, but there is no clearly-discernible atomic structure. The symmetry of the output is readily apparent, however, with octagonal patterns forming on the faces of the unit cell. Whilst this lack of learning may just be a limitation of the number of training structures used, this result suggests that 3D-Unet is learning some of the underlying 3D symmetry present by comparing equivalent diffraction peaks.

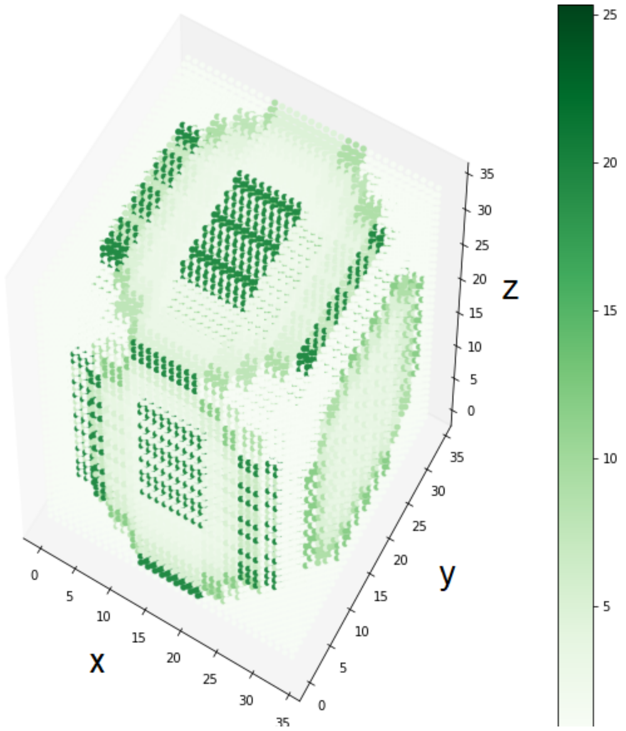


Figure 4: Output from 3D-unet model showing voxels which exhibit symmetry-related magnitudes.

By visualising only those voxels with a magnitude greater than 70% of the maximum voxel value (Fig. 5) it is clear that the voxels with maximum value do not correspond with the underlying atomic structure (shown in red). Instead, they appear in a single plane perpendicular to z, with a spacing similar to the horizontal spacing between atoms in the ground truth. This might suggest that training with a greater number of samples and/or a modified UNet architecture, could lead to a better overall representation.

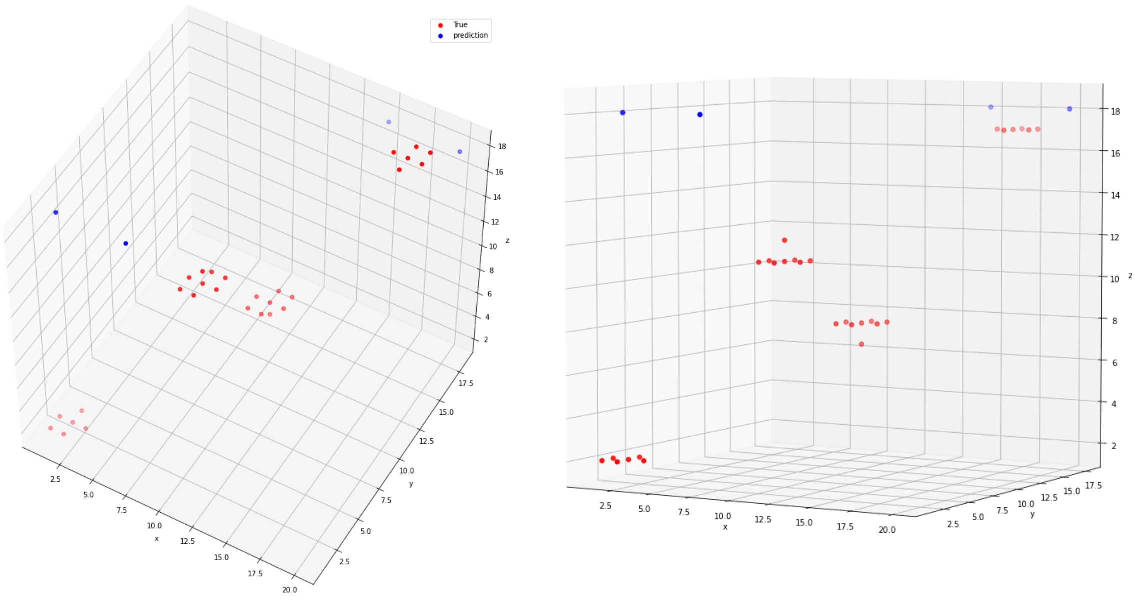


Figure 5: Two different orientations showing the predicted fluorine positions (with more than 70% of the maximum voxel value) from the 3D-unet model, with ground-truth positions shown in red.

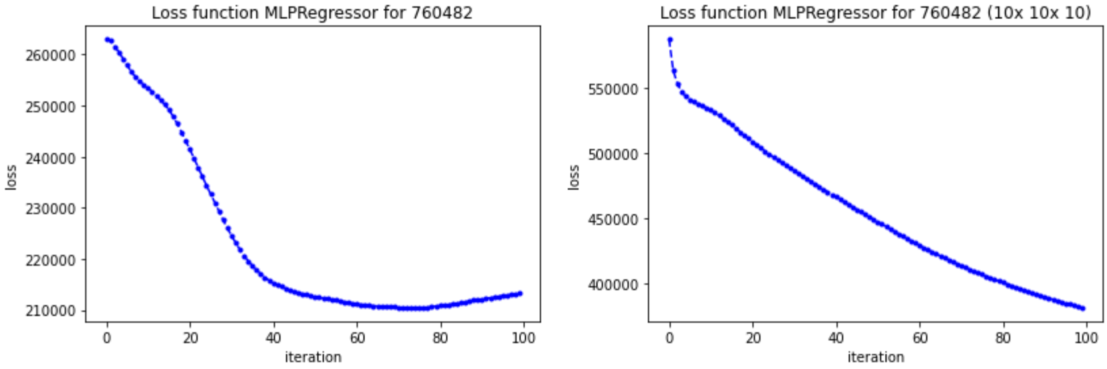
6.2 MLP Regressor

MLP Regressor showed more accurate predictions than 3D-Unet, and a greater ability to resolve centres of atomic density. The variation in loss with training iteration is a good indicator of the training progress of the network. The curve for 20 x 20 x 20 inputs (Fig. 6a) shows the expected shape; a rapid decrease in loss leading into a much slower decrease tending towards a plateau. However, the loss function for the 10 x 10 x 10 inputs does not show the plateau being reached over the same training set, and the absolute value of the loss is more than twice that of MLP20 (Fig. 6b). MLP10 is not achieving convergence at the same rate as MLP20, and the lack of convergence in MLP10 is illustrated in the lower accuracy of the predictions compared to that for the 20 x 20 x 20 training inputs. Increasing the training time or modifying the learning rate from the default may allow MLP10 to converge, but it is unclear whether the overall loss will approach that of MLP20.

6.2.1 MLP20 Predictions

An example prediction using the 20 x 20 x 20 diffraction input is shown in Fig. 7. Whilst most voxels show some degree of weak activation, there are regions showing greater intensity such as the points centred around (17, 17, 10). By filtering only those voxels showing at least 70% of the maximum voxel value (which occurs at (0,0,0)) this becomes more clear (Fig. 8).

Comparing the predicted regions with the correct distribution, it is clear that although the distributions are not identical, the model has correctly achieved four clusters of high probability (with one centred at the origin). Comparing these four clusters, there is some evidence that the



(a) MLP20

(b) MLP10

Figure 6: Loss curves for the training of MLP Regressor across 100 iterations of the training set for (a) 20 x 20 x 20; and (b) 10 x 10 x 10 inputs

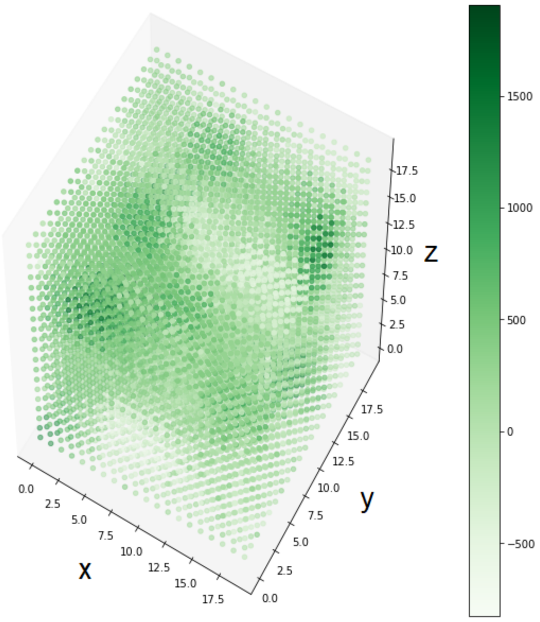


Figure 7: Fluorine crystal prediction from the MLP20 model. The plot shows the total map of electron density distributed over the voxelised unit cell. The clustering of higher electron density is shown in dark green.

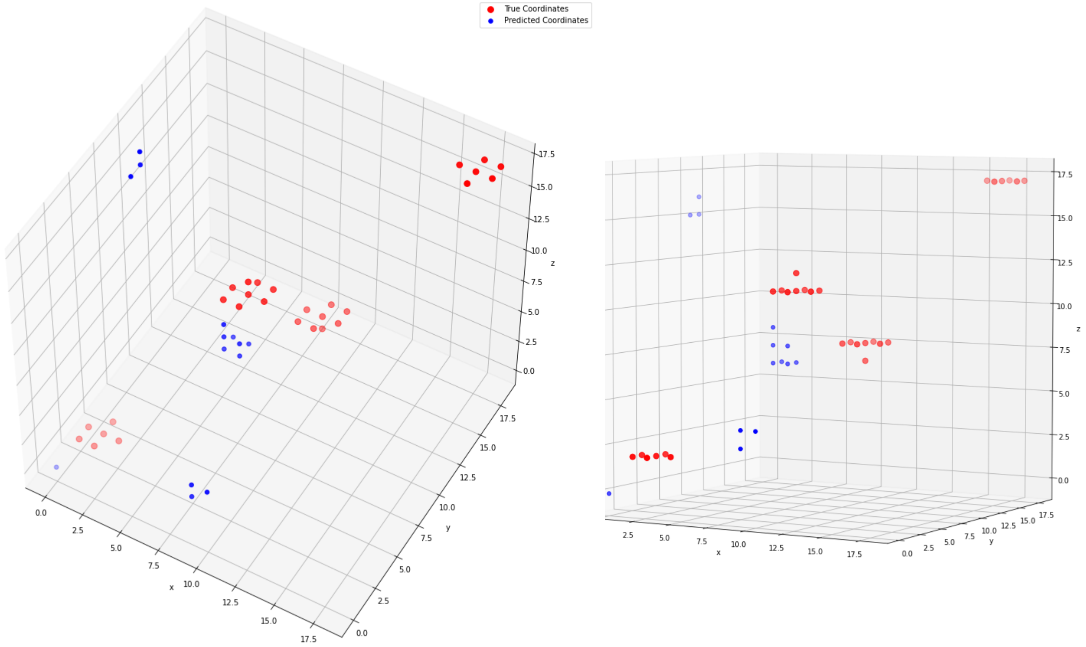


Figure 8: Plots showing two views of the peaks in electron density filtered by points above 70% of the magnitude of the maximum electron density in the voxelised structure. The blue points show the predicted points of high intensity and the red are the true points.

distances and angles between maxima share similarities, although their relative orientation/or location does not match the true structure.

Given that the phase information is lost from the diffraction signal, it is perhaps not surprising that translational differences between the input and predicted structures may occur. For this reason, vector differences between maxima were calculated so that translational similarities could be determined. Table 1 shows the vector distances (in voxel units) between pairs of atoms in the test structure, and the nearest corresponding distance between maxima in the predicted structure (calculated from the most intense 80 voxels). Clearly, the model predicts significant

Table 1: Vector shifts between atoms in the true structure, and corresponding vectors occurring in the predicted structure. The order of atom labels (a–d) are arbitrary, and vectors are given in voxel units.

Ground truth labels	Ground truth vector $\Delta(x, y, z)$	Prediction vector $\Delta(x, y, z)$
(a, b)	(6,6,9)	(7,7,8)
(a, c)	(9,9,6)	(8,8,7)
(a, d)	(15,15,15)	(15,15,15)
(b, d)	(9,9,6)	(8,8,7)
(c, d)	(6,6,9)	(7,7,8)

atom density at a similar separation as the true model, although the absolute location may be different. Work is on-going to include this translational flexibility in the model training, such that the MLP is not penalised for suggesting a model with the incorrect origin (as this can often be variable for the same structure).

6.2.2 MLP10 Predictions

Compared to the MLP20 model, the predictions from the MLP10 model show much less spread of atomic density; figure 9 plots the predicted voxel distribution using the same 70% cut-off as the MLP20 data, which reveals that much of the atom density is localised within a single voxel. Whilst this could link with the incomplete training as seen in fig. 6b, it may also relate to the relatively coarse 10 x 10 x 10 binning of the ground truth data. Whereas within MLP20 each atom occupies multiple voxels due to the Gaussian smearing applied, in the MLP10 model each atom only occupies a small number of larger voxels. As such, during model training it is much harder for the loss to be minimised as any slight deviation in prediction will miss the 'True' atom position. In contrast, a slight discrepancy in the MLP20 model is still likely to align with the tail of a Gaussian distribution, systematically improving the model. Work is on-going to explore different loss functions more suited to judging the proximity of probability distributions, which may solve this issue.

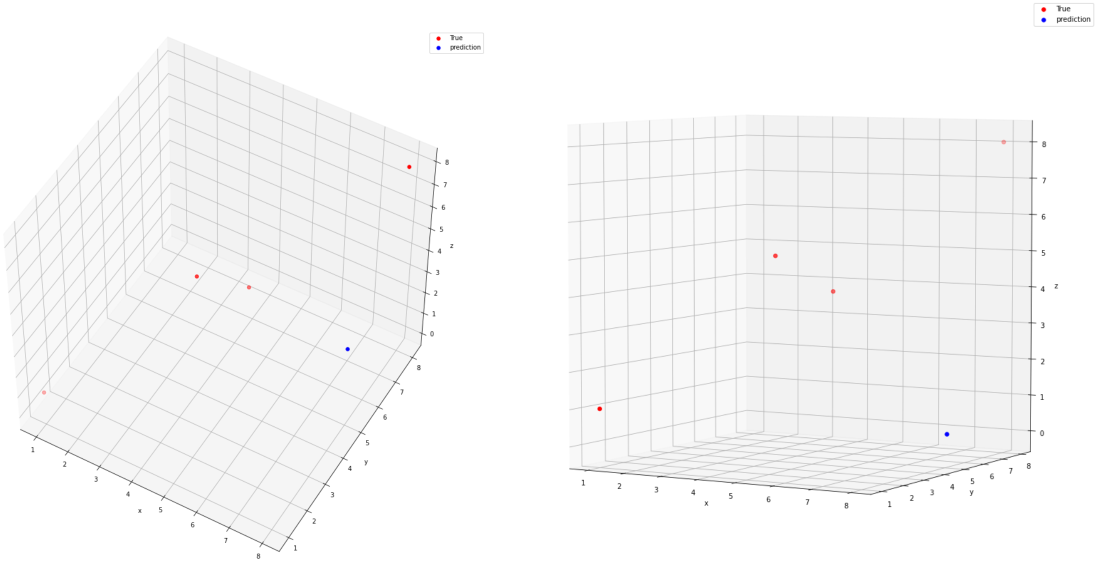


Figure 9: Plots showing the peaks in electron density filtered by points above 70% of the magnitude of the maximum electron density in the voxelised structure. The blue points show the predicted points of high intensity and the red are the true points.

7 Conclusions & Future Work

Neural networks (both convolutional and fully-connected) have been applied to the problem of resolving atomic structures from diffraction data, and both have shown some ability to correctly assign atomic positions. While the CNN model chosen (3D-Unet) proved unsuitable for learning atomic positions, there is strong evidence that the network is learning some important aspects of crystallography, such as the underlying symmetry. This prediction could potentially be improved by using larger sets of training data, however this will require specialised computer architecture due to the large memory requirements. Alternatively, a simpler CNN architecture could be developed to allow the processing of more data, however this was beyond the scope of this project.

As a simpler alternative to 3D-Unet, a multi-layer perceptron (MLP) model was trained on the full data set of 12,000 structures. This model is able to produce the expected regions of high atomic density, resembling realistic atom positions. Additionally, the vector relationships between these regions match those of the underlying test structure, although the actual locations of the maxima are shifted from the ground truth. Work is on-going to expand upon this vector-based comparison, and also to implement the effect of periodic boundary conditions when determining training loss. Training on smaller input data appears to be slower due to the reduced 'spread' of atoms across voxels, but clearly training on smaller data has advantages in terms of computer memory etc. Future work will look at how this training can be improved further.

8 Outputs, Data & Software Links

A poster of this work was presented at the AI^{3}SDSummer Intern conference by SJS. The project was also presented as part of an Afton Chemicals-funded event at the University of Edinburgh for summer intern students (SJS).

Code developed during this project is currently stored on a private repository within the Functional Materials Group GitLab, but will be made available prior to publication of the results.

9 References

References

[1] Özgün Çiçek, Ahmed Abdulkadir, Soeren S. Lienkamp, Thomas Brox, and Olaf Ronneberger. 3d u-net: Learning dense volumetric segmentation from sparse annotation. *CoRR*, abs/1606.06650, 2016.