Quantifying the resolution of X-ray diffraction images through the usage of deep learning techniques

Aleksander Cichosz[1], Vardan Martirosyan[1], Ryan Stofer[1], Teo Zeng[1], Robin Liu[1], Derek Mendez[2]

[1] University of California, Santa Barbara; [2] Stanford SLAC National Accelerator Laboratory

UC SANTA BARBARA | Data Science Initiative

Abstract

Rising data rates in protein crystallography pose a challenge for experimentalists. This poster presents multiple Convolutional Neural Networks (CNNs) created to automate analysis of images generated through X-ray crystallography. With modern advancements, thousands of diffraction images are now being created per second, and crystallographers must analyze these images just as fast to optimize experimental resources. We focused on two aspects of diffraction image characterization usually done by hand. First, a classification model was trained to categorize images as single- or multi-lattice, as multi-lattice images are ill-suited for future analysis. Second, a regression model was used to estimate the resolution of each image. Both models follow a ResNet[1] architecture, and were trained using simulated images. Our classification CNN achieved a 94% accuracy and our regression CNN achieved a 0.96 Pearson correlation value on simulated data. The models also performed qualitatively well on experimental data.

Introduction

Proteins play a fundamental role in biological processes, and understanding their structure and function is crucial for advancements in fields such as medicine, drug development, and bioengineering. X-ray crystallography, a powerful technique, provides invaluable insights into protein structures by capturing diffraction patterns generated when X-rays pass through protein crystals.

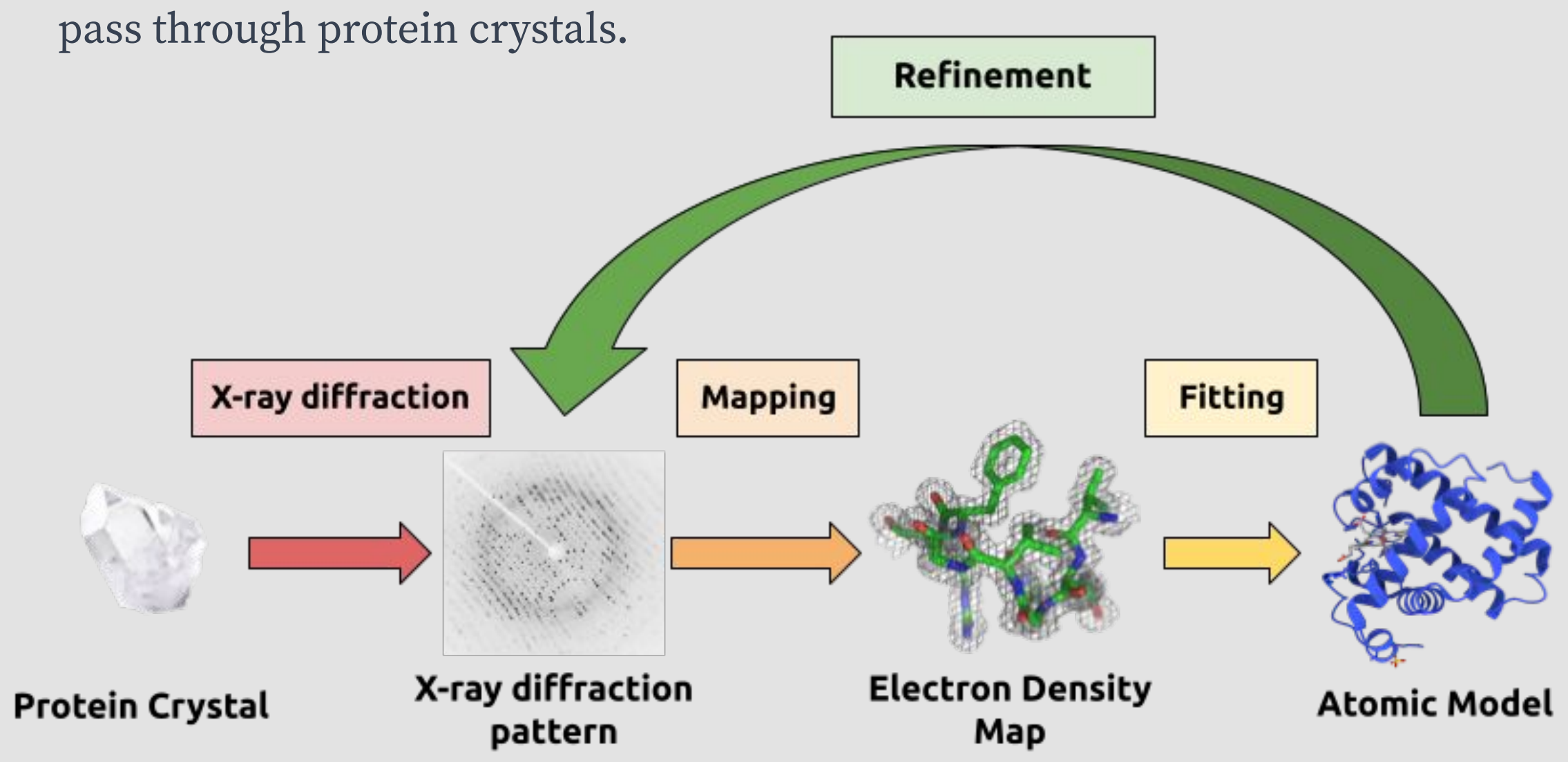

Fig. 1: General workflow process of X-ray crystallography for protein reconstruction.

In collaboration with SLAC SSRL, our team is undertaking an innovative project to streamline the analysis of images obtained through X-ray crystallography. By harnessing the capabilities of Convolutional Neural Networks (CNNs), we aim to develop a versatile and efficient multi-task deep learning model. This model will enable laboratories worldwide to expedite the process of classifying usable X-ray diffraction shots and their respective resolution which is used to determine a protein's electron density map.

Our project is divided into two teams, both focused on a specific aspect of the multi-task learning model development. The multi-lattice group developed a CNN that distinguishes between single- and multi-lattice shots. The resolution group constructed a CNN to predict the resolution of crystal structures, a critical parameter for protein reconstruction.

Key Terms

- **Shot**: A singular diffraction image from one crystal.
- **Resolution**: Measure of the finest details or features that can be distinguished in a crystal structure, indicating level of structural clarity. Used for generation of a protein's electron density map.
- **Single-Lattice**: A shot representing diffraction from a single crystal lattice.
- **Multi-Lattice**: A shot representing diffraction from multiple crystal lattices.

Datasets

Our datasets consisted of X-ray diffraction shots in two different forms: **simulated images** (for training) and **real user data** (for testing and evaluation). The variables that were measured were the images themselves, and the data dimensions were 512 x 512 pixel arrays. The datasets we used are listed below:

Resolution Datasets:

- RESMOS2: 10,000 labelled simulated diffraction shots.
- Master: 45,000 labelled simulated diffraction shots.
- MORE_DATA: 26 real user datasets of varying resolutions (1.42Å to 5.45Å) each containing a range of either 500, 700, or 1800 shots.

Multi-Lattice Datasets:

- Master_baxter: 100,000 labelled simulated diffraction images. Generated using a Gaussian distribution of crystal orientations.
- Uniform Dataset: 50,000 labelled simulated diffraction images. Generated using a uniform distribution of crystal orientations.
- Test.baxter: 512 unlabelled real user data of single-/multi-lattice shots.

Methodology

All our models were Convolutional Neural Networks (CNN). CNNs are a class of neural networks that specialize in processing data that has a grid-like topology, such as image data. All CNNs have three key layers: a convolutional layer, a pooling layer, and a fully connected layer.

- **Convolutional layer**: Applies filters to extract important patterns, allowing the network to recognize complex features and objects.
- **Pooling layer**: Reduces the spatial dimensions of feature maps, preserving important features while reducing computational complexity and enhancing translational invariance.
- **Fully connected (FC) layer**: Connects every neuron from the previous layer to every neuron in its layer, enabling high-level feature extraction and classification.

Our CNN models implemented are based on ResNet, a cost-effective deep learning algorithm used for image recognition. ResNet introduces **skip connections** that directly link earlier layers with deeper ones, allowing information to bypass certain layers. Classification models for multi-lattice detection used ResNet34 whereas regression models for resolution prediction used ResNet18 and ResNet50.

Note: The ResNet number represents the number of layers in the model.

ResNet Model Architecture

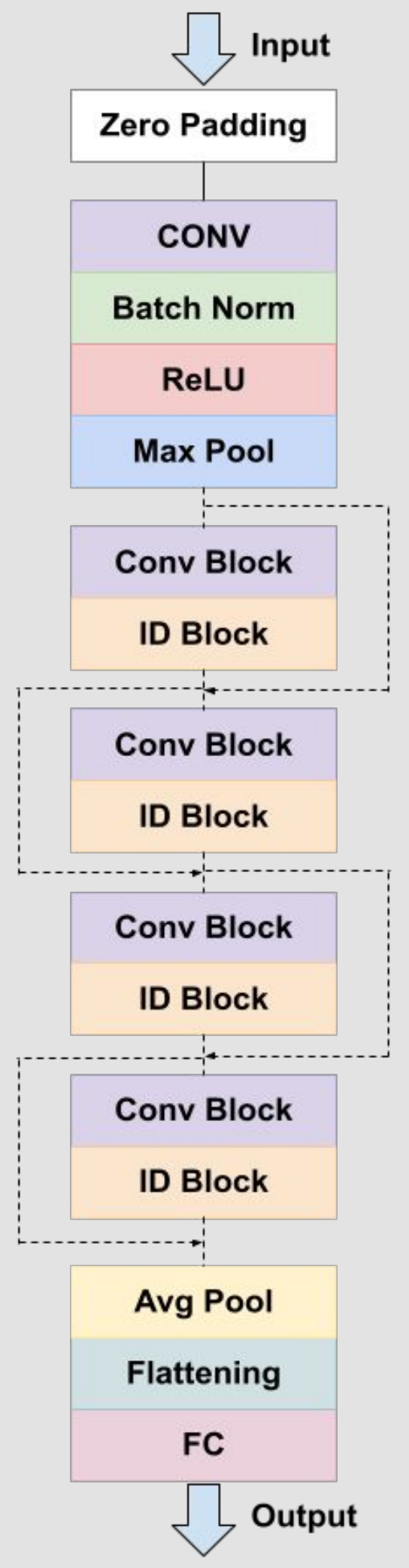

Fig 2: General CNN architecture framework for a ResNet Model

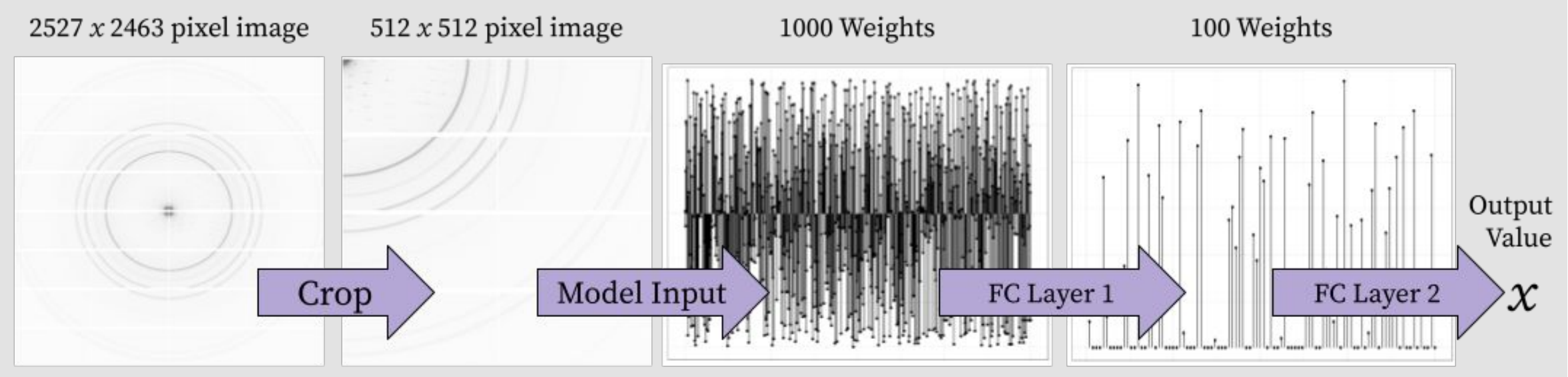

Fig 3: Flowchart showing the downsampling and modeling process of an X-ray diffraction shot image.

Difference Between a Multi and Single Shot X-Ray Crystallography Image

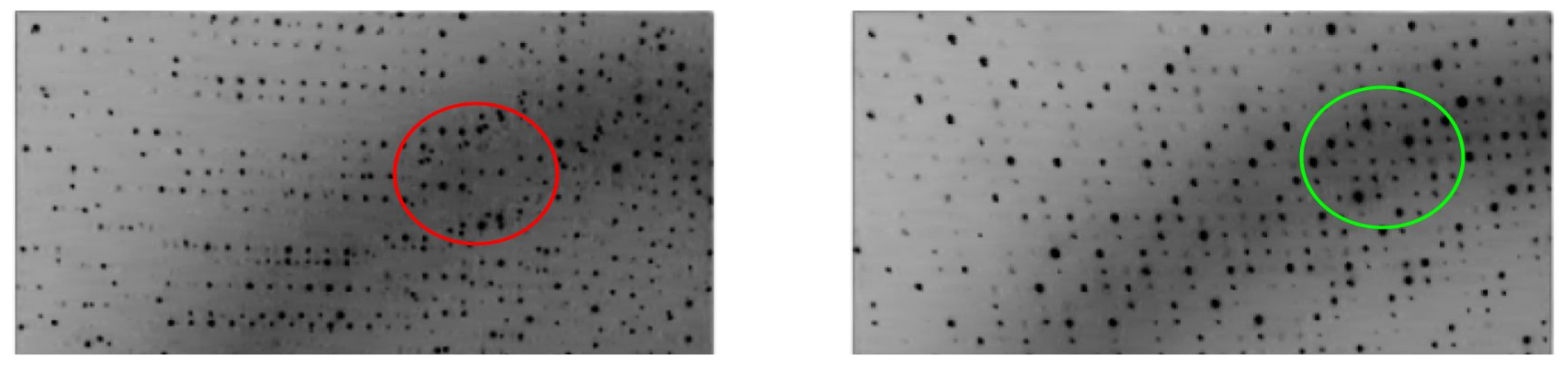

Fig 4: Difference between a single-lattice shot and a multi-lattice shot. On the left hand side in the red circle, we can see disordered scattering resulting from multi-lattice diffraction. In the green circle on the right, we can see cleaner diffraction indicative of single-lattice scattering. One can normally tell that a shot is multi-lattice if the peaks (Bragg reflections) in the image are grouped very closely to one another in a non-ordered fashion, as illustrated in the left figure.

Results: Multi-Lattice Detection

Model Training Performances for Multi-Lattice

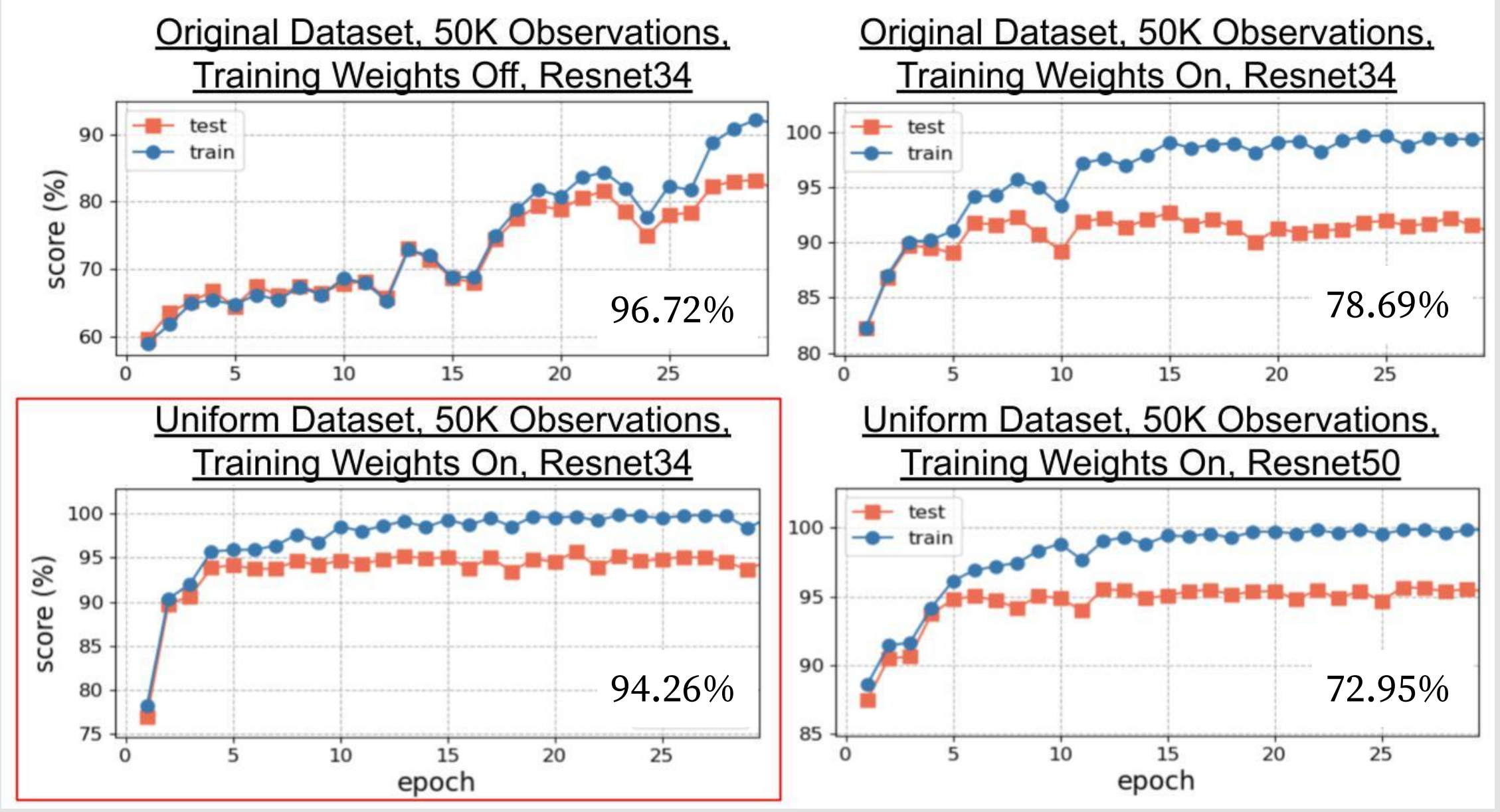

Fig 5: Four graphs representing the performances of the four of the six models that we made in the multi-lattice group. Each graph represents the accuracy score for each model on simulated data. Additionally, we tested the four models on hand selected experimental images that had been chosen based on having single-lattice characteristics. Our model's accuracy scores on those images are shown on the bottom right of each subplot above. Additionally, the overall best model is boxed in a red square.

Results: Resolution Prediction

Best Model's Prediction vs Ground Truth on Training Data

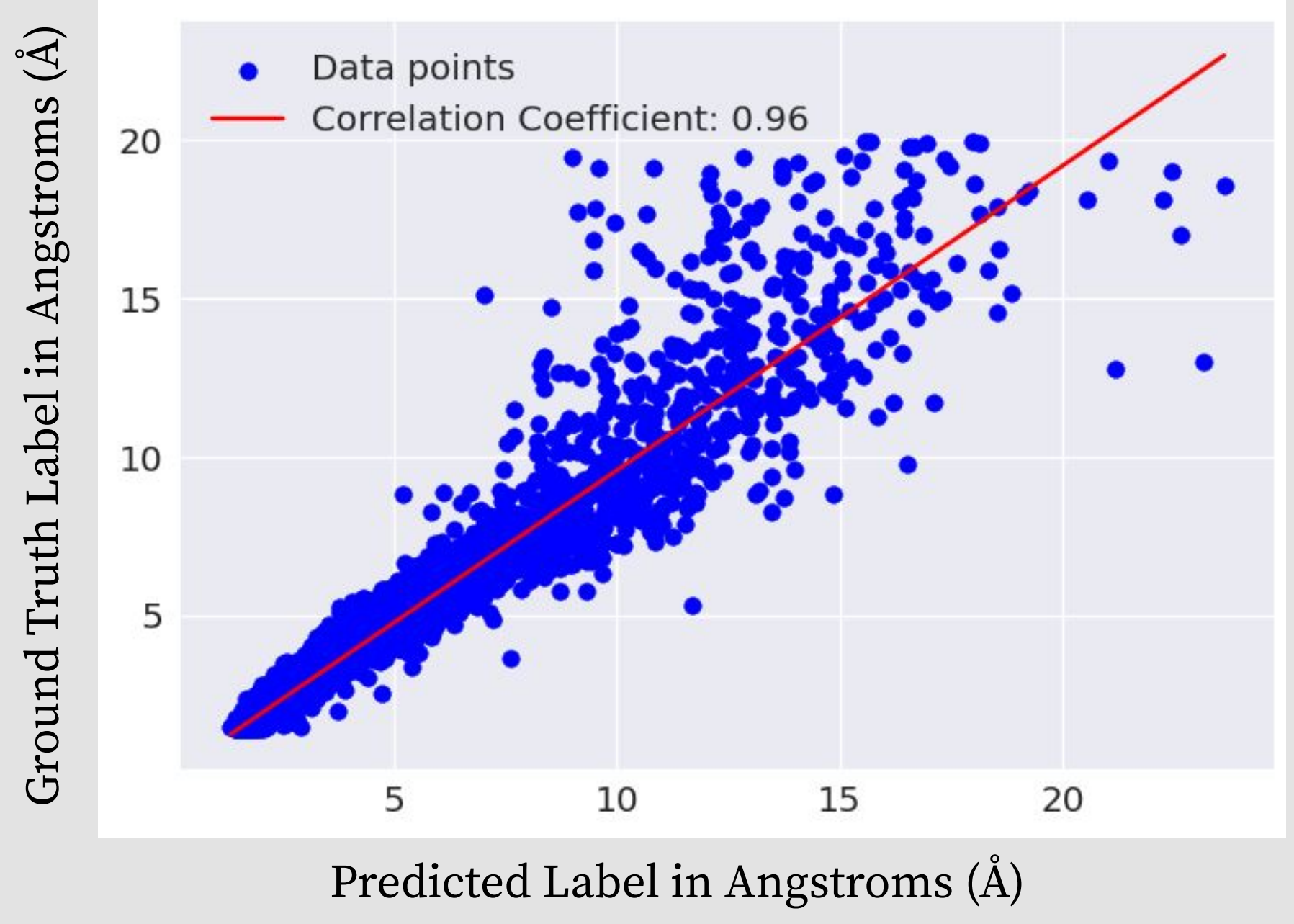

Fig 6: Best model's prediction vs ground truth on simulated training data.

Mean and Highest Prediction Compared to Ground Truth Label

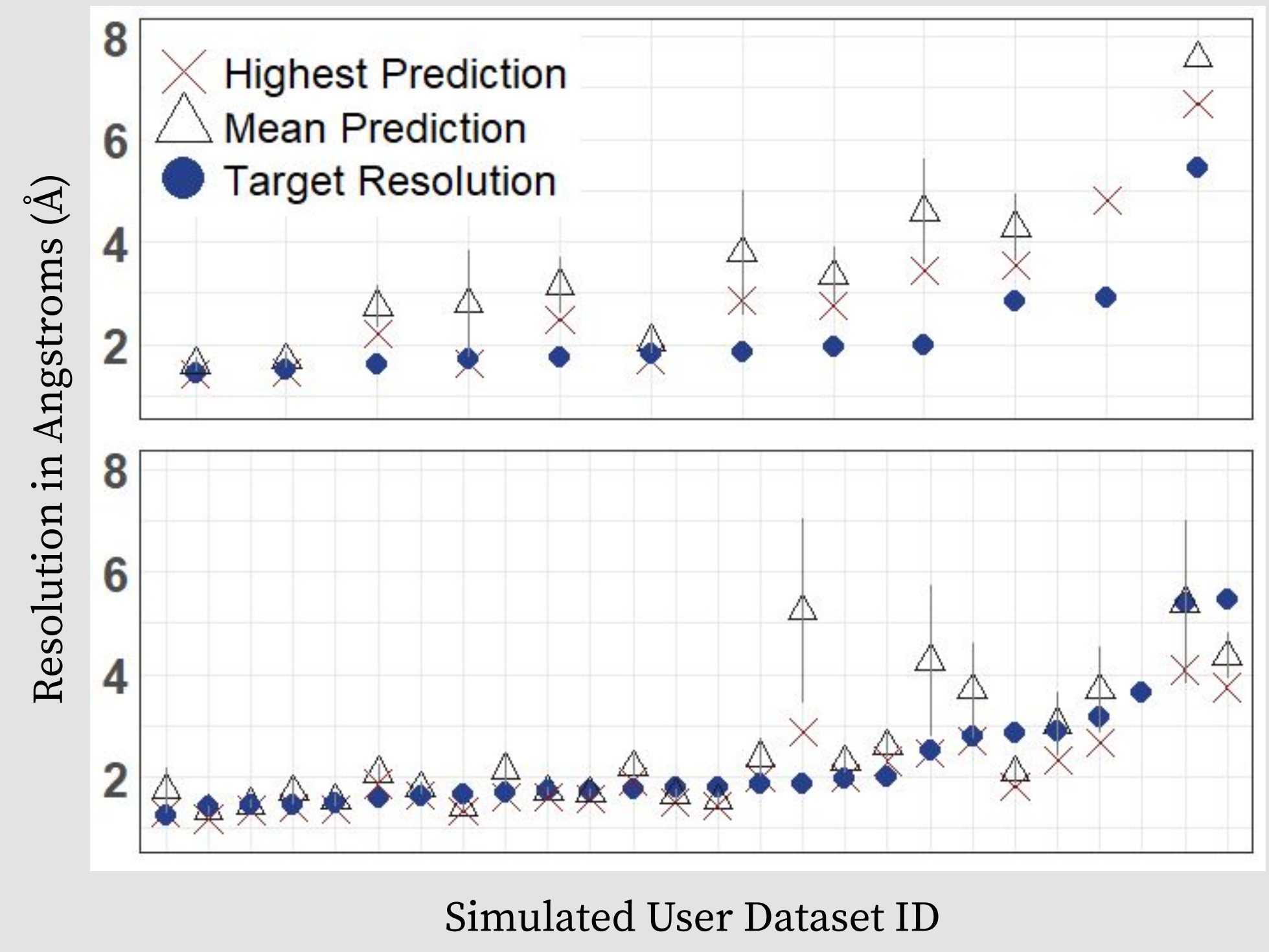

Fig 7: Model prediction compared to true label on simulated experimental user datasets. We see from the plot that a majority of the predictions lie close to the ground truth, although the accuracy wanes as the target resolution increases. ***Note: In resolution terms, the finest (or highest) resolution quality images correspond to the lowest resolution number.***

Summary of findings

Multi-lattice:

- The multi-lattice model that was trained on the uniform dataset, with training weights on, and with Resnet34, performed the best out of all of our other models with a testing accuracy of **94%**.
- Most of the models performed quite well, with all of them reaching at least an **80%** testing accuracy score.

Resolution:

- Most Resnet18 and Resnet50 models achieved a high Pearson correlation coefficient of around **0.96** when quantifying the resolution of simulated X-ray diffraction images during training.
- Our best model evaluated more than **75%** correctly out of the total 26 simulated user data sets within 2 standard deviations of the target resolution.

General Remarks:

- Our best models achieved qualitatively good results for both tasks, signifying that deep learning models are indeed capable of increasing the speed and scalability of X-ray crystallography analysis.
- By simulating training data, we easily compiled vast and diverse training datasets. However, since the models were trained on simulated images alone, it may be challenging for them to properly handle certain image artifacts resulting from e.g., atypical salt and ice scattering and background signals in real diffraction shots.

Future work

- If allotted more time, we would like to extend our work further by testing our model on more real user data and assessing its performance on both tasks.
- We are currently in the process of submitting our work as a manuscript for publication to *Acta Crystallographica Section D*, a journal centered around articles covering structural biology.
- We plan on deploying our model to be used by researchers at the SLAC SSRL to streamline the process of identifying clean X-ray diffraction shots and classifying their respective resolution with the vision of possibly making our model open source for all curious X-ray researchers to use.

References and Acknowledgments

We would like to thank SLAC and the Structural and Molecular Biology (SMB) team at SSRL for sponsoring this project and special thanks to Trevor Ruiz, Derek Mendez, and Robin Liu for their guidance and support. We would also like to thank the California NanoSystems Institute for providing access to their computing resources.

Website link to SMB team at SSRL:
https://smb.slac.stanford.edu/

References:
[1] He, K., et al., "Deep Residual Learning for Image Recognition," arXiv: 1512.03385 [cs.CV], Dec. 2015.

Github Repository: