XPSI: X-ray Free Electron Laser-based Protein Structure Identifier

Paula Olaya
University of Tennessee, Knoxville, USA
polaya@vols.utk.edu

Michael R. Wyatt II
University of Tennessee, Knoxville, USA
mwyatt@udel.edu

Silvina Caíno-Lores
University of Tennessee, Knoxville, USA
scainolo@utk.edu

Florence Tama
RIKEN, Japan
florence.tama@nagoya-u.jp

Osamu Miyashita
RIKEN, Japan
osamu.miyashita@riken.jp

Piotr Luszczek
University of Tennessee, Knoxville, USA
luszczek@icl.utk.edu

Michela Taufer
University of Tennessee, Knoxville, USA
mtaufer@utk.edu

***Abstract*—A protein's structure determines its function. Different proteins have different structures; proteins in the same family share similar substructures and thus may share similar functions. Additionally, one protein may exhibit several structural states, also named conformations. Identifying different proteins and their conformations can help solve problems such as determining the cause of diseases and designing drugs. X-ray Free Electron Laser (XFEL) beams are used to create diffraction patterns (images) that can reveal protein structure and function. The translation from diffraction patterns in the XFEL images to protein structures and functionalities is non-trivial. In this poster we present the first steps into the design and assessment of a software framework for the identification of XFEL images. We call the framework XPSI (XFEL-based Protein Structure Identifier). We quantify the identification accuracy and performance of XPSI for protein diffraction imaging datasets including different protein orientations and conformations with realistic noise incorporated.**

1. Introduction

Proteins and other biological molecules are responsible for many vital cellular functions. The structure of the protein determines its functionality. Identifying the information of a protein structure is helpful to understand the protein functional mechanisms, which can help solve many difficult problems such as determining the cause of diseases and designing drugs. The X-ray Free Electron Laser (XFEL) provides beams that are applied to proteins and generate diffraction patterns (images) that can reveal the inner protein structure [1]. Specifically, three properties can be embedded in an image: the orientations of a protein conformation, the conformations of a folded protein, and the different conformations of different proteins.

In this work we focus on identifying the first two properties: orientation and conformation. Orientation refers to the placement of the incident beam with respect to a protein structure and is defined by the three Euler angles: Φ (Azimuth), Θ (Altitude), and Ψ (Rotation angle). Conformation determines the overall shape of the molecule. We present the design, implementation, and validation of a software framework for the predictions of proteins' orientation and conformation embedded in XFEL images. We call this framework XPSI (XFEL-based Protein Structure Identifier).

The contributions of this work are as follows:

- The prototype of XPSI, a framework to predict structural properties such as conformation orientations and structural conformations from datasets of high and low energy X-ray diffraction patterns.
- The quantification of the prediction accuracy and performance of XPSI for datasets representing realistic scenarios for protein diffraction imaging (e.g., different orientations and conformations with noise).

2. Framework Overview

Figure 1 presents the XPSI framework. The input data are the diffraction patterns (i.e., images that embed the structure of proteins). The diffraction patterns are generated by simulations or experiments using an X-ray free-electron laser (XFEL) beam. The higher the beam intensity, the higher the resolution and precision in diffraction patterns. Patterns are processed by an autoencoder that captures key information and produces a tensor representation of each pattern. The autoencoder consists of an encoder and a decoder. The encoder has 3 convolutional filters and down-sampling layers. The decoder has the reverse structure of the encoder. The new latent space is used to train and validate traditional machine learning models such as k-nearest neighbors (kNN). We use a kNN-angle regressor for predicting the orientation and a kNN-classificator for predicting different protein conformations.

3. Dataset, Predictions, and Results

Dataset and Tests We quantify the prediction accuracy and performance of XPSI by identifying the structural

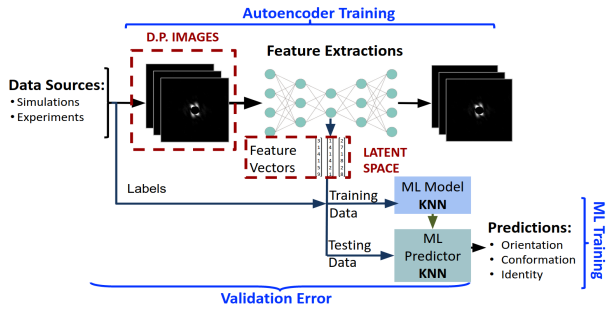

Figure 1: Framework components.

properties (both orientation and conformation) of diffraction patterns of a Eukaryotic Elongation Factor 2 (eEF2) protein. Our dataset consists of two conformations, PDB: 1n0u and 1n0vc, each with 39,692 different orientations. To have a more realistic scenario we use the images generated by two beam intensities: high intensity and low intensity. The patterns produced by the low intensity beam have a worse resolution, thus they are noisier and more realistic. We define two test cases:

- **Test 1: Prediction of orientations.** For 1n0u (39,692 samples), we predict two Euler angles (azimuth Φ and altitude Θ) defining a protein orientation.
- **Test 2: Prediction of orientations and conformations.** For 1n0u and 1n0vc, with 39,692 samples each, we predict two Euler angles (azimuth Φ and altitude Θ) and the conformation between the two.

Orientation and Conformation Metrics. To measure accuracy, we define two error metrics: the orientation's *error degree* and *conformation accuracy*. We use the Haversine formula to calculate the *error degree* and determine the distance in degrees of two points on a sphere given Φ (azimuth) and Θ (altitude). The number of correct predictions (both true positives and true negatives) over the total number of predictions defines the *conformation accuracy*.

Platform. Due to the amount of data and the need for an effective framework, high performance resources are required for executing and identifying the proteins with XPSI. We execute our tests on a single node Power9 (128 GB RAM) with 2 GPUs –NVIDIA Volta V100, from the Tellico cluster. We measure the training cost of the autoencoder as well as the training and validation cost of the machine learning model (kNN). All measurements are the average over 20 runs.

Autoencoder Training. In Table 1, we present the average training time from the autoencoder. The autoencoder training time is proportional to the number of images used as input. For the first test where we only have 39692 samples, it takes 45 mins to train. For the second where we have double the amount of images, we also double the training time.

Machine Learning Training. In Table 2, we present the time measurements over 20 trials for the kNN training and validation. The times are proportional to the number of samples.

Test	Prediction	Data Size	Time [s]
1	Orientation $[\Phi, \Omega]$	39692	2790.8
2	Orientation + Conf.$[\Phi, \Omega$,conf]	79384	5415.6

Table 1: Autoencoder time performance metrics.

Test	Prediction	Data Size	Training Time [s] (90%)	Validation Time [s] (10%)
1	Orientation	39692	0.07	0.10
2	Orientation + Conf.	79384	0.34	0.66

Table 2: Machine Learning time performance metrics.

Accuracy Results. Empirically, we select the k in the kNN method that allows the lowest validation error for each test. We observe that for Test 1, where we only predict orientation, 99% of the data samples have an error degree within 2.2° for both high and low beam intensities. The other 1% for the high intensity have an error degree within 4°, while for the low intensity we get 5°. This *error degree* is negligible and does not affect the scientific interpretation of the protein structure. For Test 2, where we augment to predict orientation and conformation, we have the same error degree as in Test 1, and 100% of accuracy when predicting the conformation for both beam intensities.

These results indicate that XPSI is a promising approach towards the identification of protein structures with respect to its computational requirements. XPSI provides high prediction accuracy of protein properties such as orientation and conformation with a total computational time cost of approximately 1.5 hours.

4. Conclusions

We design, implement, and evaluate XPSI (X-ray Free Electron Laser-based Protein Structure Identifier), a framework capable of predicting two structural properties of proteins from their diffraction patterns. Our framework predicts orientation with an error degree within 2.2° and identifies the conformation from two different datasets (i.e., 1n0u and 1n0vc) with an accuracy of 100% for the eEF2 protein.

Acknowledgments

This work was supported by NSF award IIS 1841758, in collaboration with the JLESC Consortium. It was also supported by JDRD Program at UTK and IBM Shared University (SUR) Award, and by FOCUS for Establishing Supercomputing Center of Excellence.

References

[1] M. Nakano, O. Miyashita, S. Jonic, C. Song, D. Nam, Y. Joti, and F. Tama, "Three-dimensional reconstruction for coherent diffraction patterns obtained by xfel," *Journal of synchrotron radiation*, vol. 24, pp. 727–737, Jul 2017. 28664878[pmid].