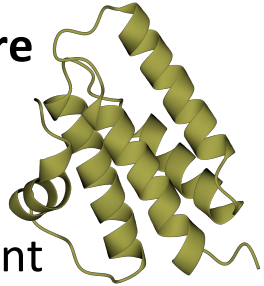
XPSI: XFEL-based Protein Structure Identifier

Paula Olaya, Mike Wyatt, Silvina Caino-Lores
Florence Tama (RIKEN), Osamu Miyashita (RIKEN),
Piotr Luszczek (ICL@UTK), and Michela Taufer(GCL@UTK)

Problem Overview

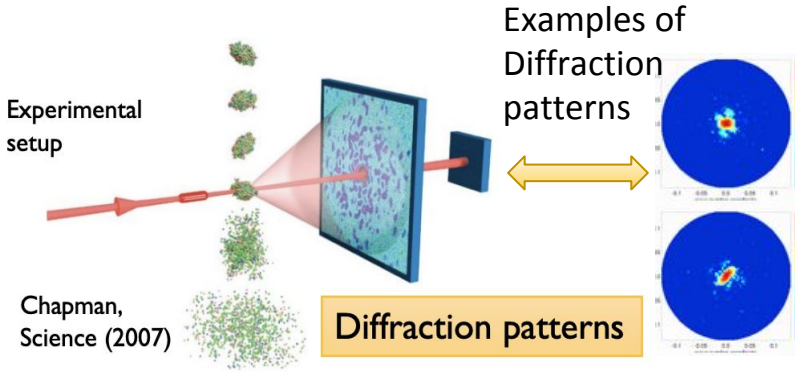
The importance of protein structure



- Protein structure determines function
- Different proteins have different structures
- One protein may exhibit several structures (conformations)
- Identifying and differentiating between protein structures is critical for:
  - Determine cause of diseases
  - Design of drugs

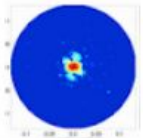
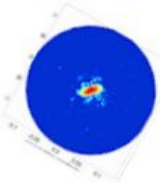
Determining structure from diffraction patterns

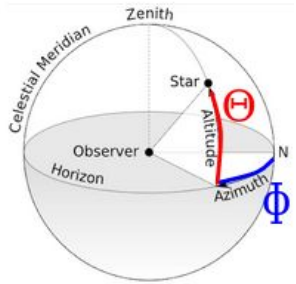
- X-ray Free Electron Laser (XFEL) beams create diffraction patterns that may reveal protein structure



Goal and Dataset

Design, implement, and validate a framework for identifying protein structural properties (i.e., orientation and conformation) from different diffraction patterns

Images	Properties	
	Orientation [ϕ, Θ]	**Conformation**
	(0°,45°)	State 1
	(15°,90°)	State 2



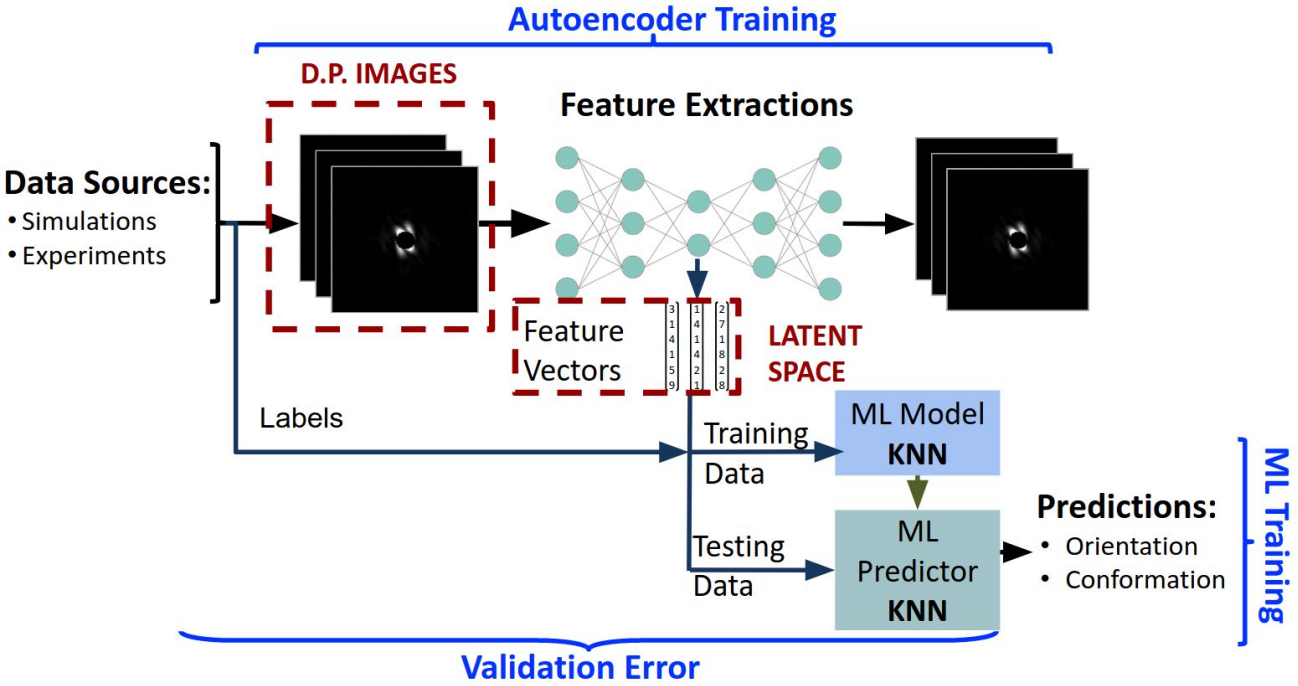
Datasets

Protein: Eukaryotic Elongation Factor 2 (eEF2)

- **39, 692 data samples** of different orientations per conformation
- **Two conformations:** 1n0u and 1n0vc
- **Two beam intensities:** High and Low

XFEL-based Protein Structure Identifier Workflow

- **XPSI predicts** structural properties such as **orientation** and **conformations**
- We **measure** the prediction accuracy and performance of XPSI



Test Cases

Test	Prediction	Data size
1	Orientation [ϕ,Θ]	39,692
2	Orientation + Conformation [ϕ,Θ, conf]	79,384

Autoencoder training time

Test	Prediction	Data size	Autoencoder training Time [mins]
1	Orientation [ϕ,Θ]	39,692	45
2	Orientation + Conformation [ϕ,Θ, conf]	79,384	90

ML training and validation time

Test	Prediction	Data size	Autoencoder training Time [mins]	kNN training time [s]	kNN validation time [s]
1	Orientation [ϕ,Θ]	39,692	45	0.07	0.10
2	Orientation + Conformation [ϕ,Θ, conf]	79,384	90	0.34	0.66

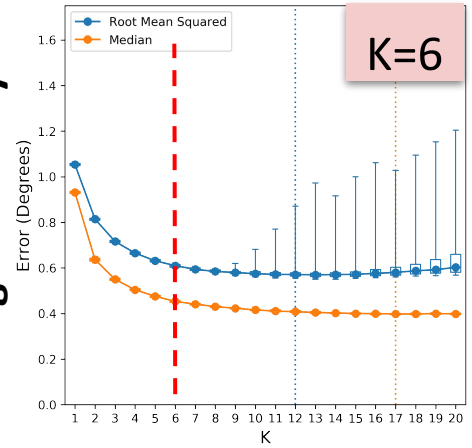
Autoencoder and ML time

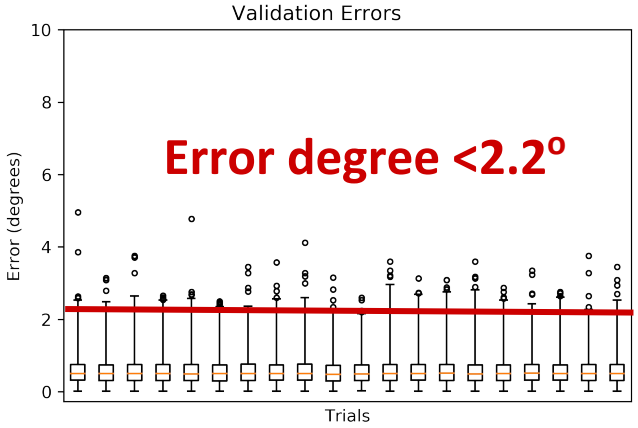
Test	Prediction	Data size	Autoencoder training Time [mins]	kNN training time [s]	kNN validation time [s]
1	Orientation [ϕ,Θ]	39,692	45	0.07	0.10
2	Orientation + Conformation [ϕ,Θ, conf]	79,384	90	0.34	0.66

- Training time proportional to processed images
- Scientists need a ~day to predict orientation using statistical models
- Our method predicts orientation + conformation in less than 2 hours

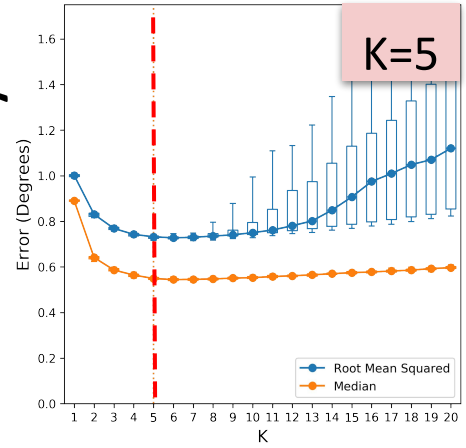
Validation Error: Orientation (Test 1)

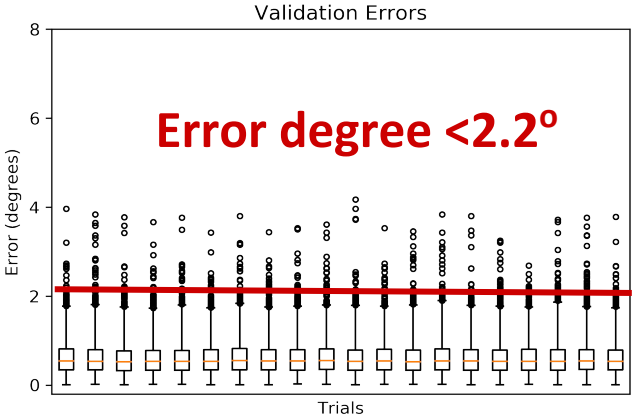
High intensity





Low intensity





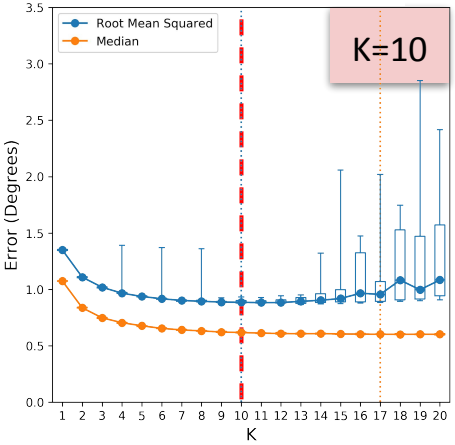
Conformation: 1n0u
Prediction = [Θ, ϕ]

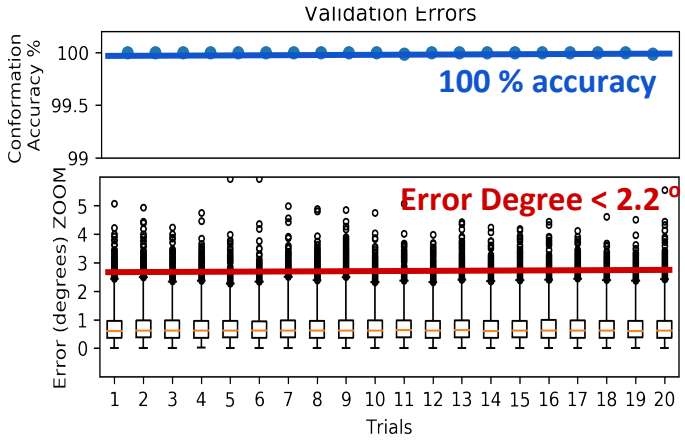
Error Degree: The distance in degrees between two points on a sphere

- An **error degree within 2.2° is negligible** and does not affect the scientific interpretation of the protein structure

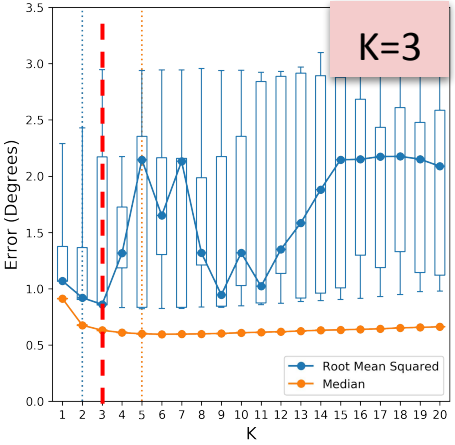
Validation Error: Orientation + Conformation (Test 2)

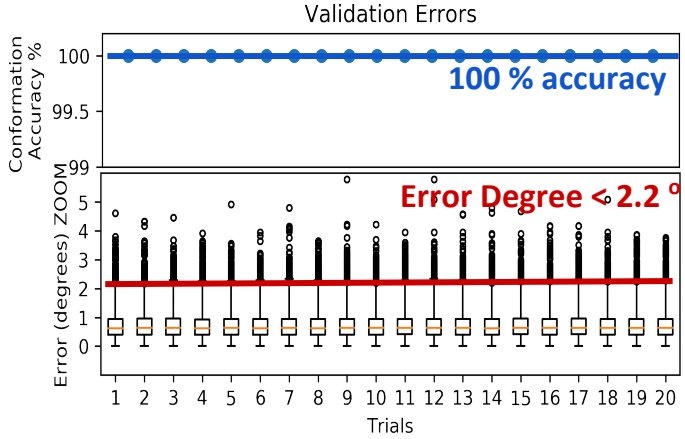
High intensity





Low intensity





Conformation: [1n0u, 1n0vc]

Prediction = [Θ, ϕ,conf.]

Conformation Accuracy:

The proportion [%] of correct predictions among the total number of cases examined

- **100% of accuracy** when identifying between the two conformations

Conclusions and Next steps

XPSI is a promising approach towards the identification of protein structures with respect to its computational requirements. Our framework predicts orientation with an error degree within 2.2^{o} and it identifies the conformation from two different datasets (i.e., 1n0u and 1n0vc) with an accuracy of 100% for the eEF2 protein

- Evaluate the generality of our framework for a broader range of datasets:
  - Intermediate conformations for the same protein (1n0u, 1n0vc, mov20, and mov53)
  - Multiple proteins with different conformations (EF2 and ribosome)
- Apply our framework to a real-world application such as, 3D reconstruction