Identifying Structural Properties of Proteins from X-ray Free Electron Laser Diffraction Patterns

Paula Olaya[*], Silvina Caino-Lores[*], Vanessa Lama[*], Ria Patel[*], Ariel Keller Rorabaugh[*], Osamu Miyashita[↡], Florence Tama[↡†], and Michela Taufer[*]

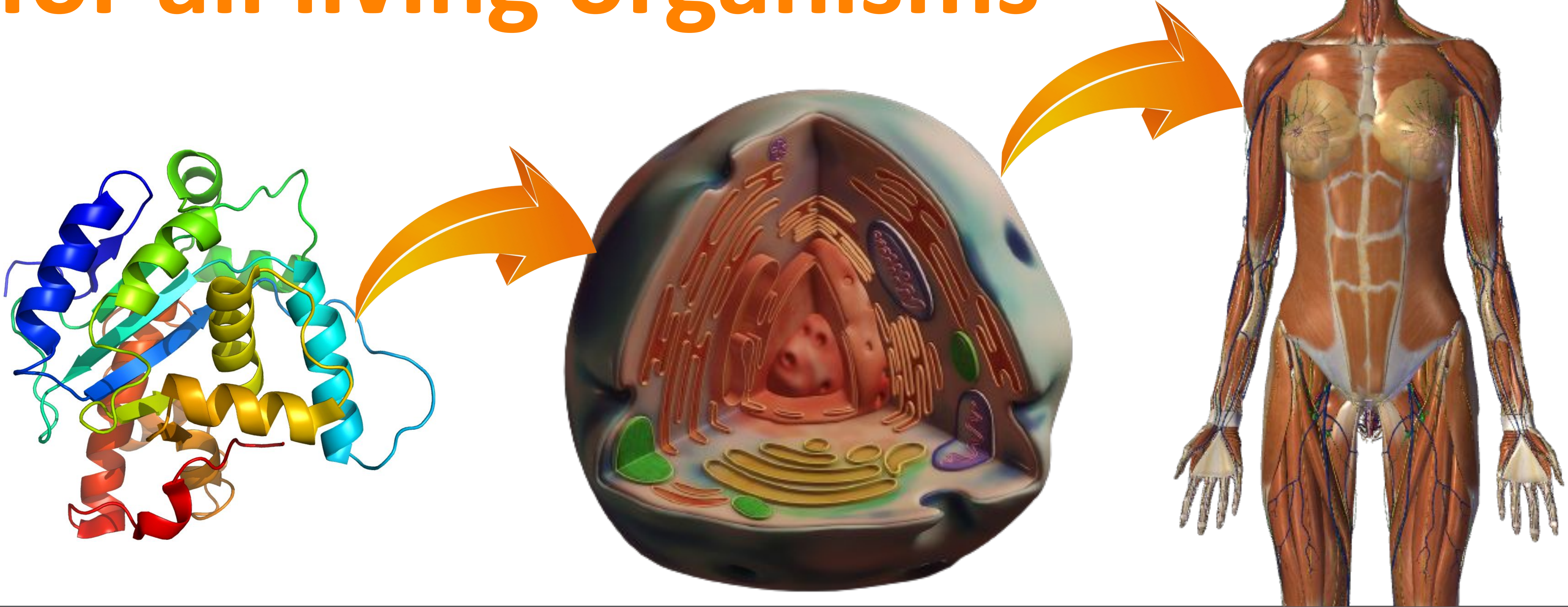
[*]University of Tennessee, Knoxville, USA

[↡]Center for Computer Science, RIKEN, Kobe, Japan

[†]Nagoya University, Nagoya, Japan

eScience 22, October 10-14, 2022, Salt Lake City, Utah, USA

Proteins are essential for all living organisms

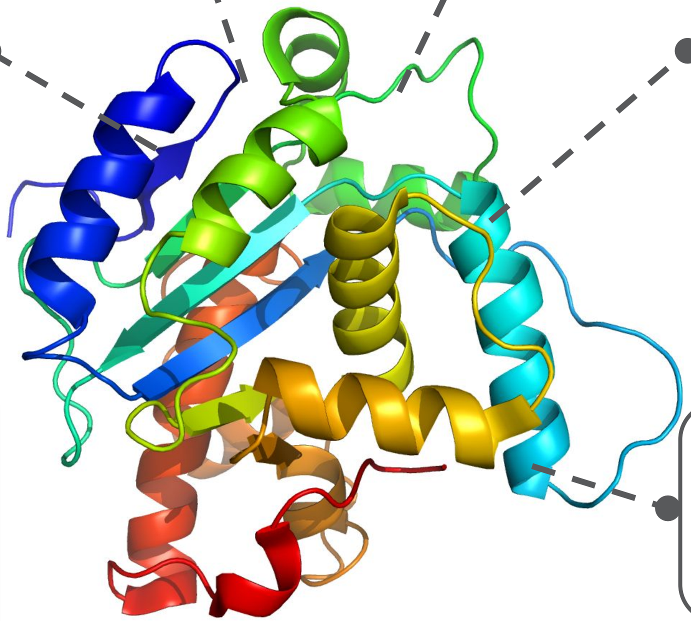


Catalysis: Increasing the rate of a chemical reaction within cells

Transportation: Moving materials within a cell and the organism

Structure: Providing structure and support for cells

Signaling: Receiving, processing, and transmitting signals within the cell and with the environment



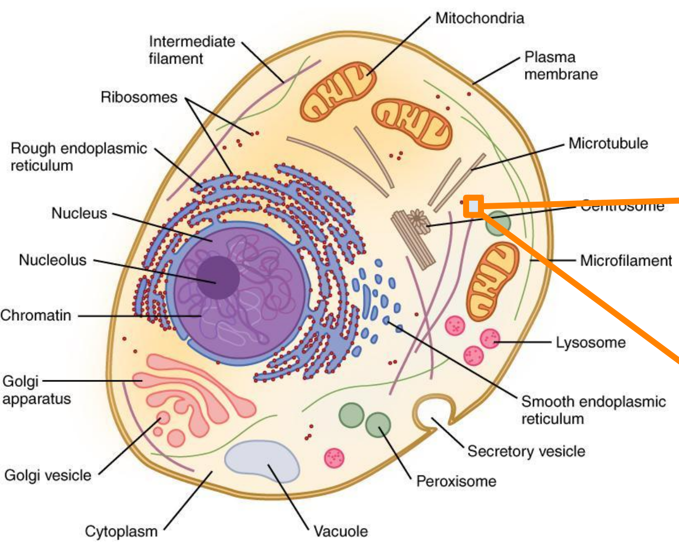
Proteins are responsible for many vital cellular functions

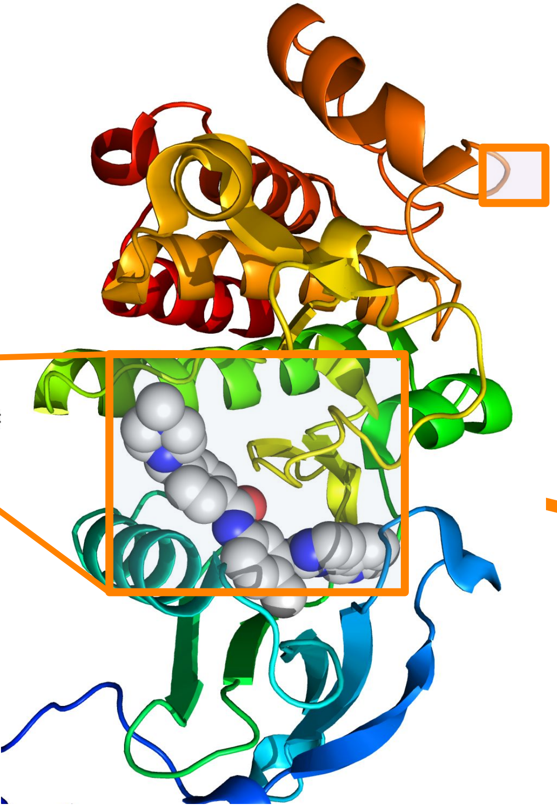
Antibodies: Helping to protect the body from foreign particles, such as viruses and bacteria

- Structural biology explains 3D structures of biomolecules
- Biomolecules are **proteins**/RNA/DNA

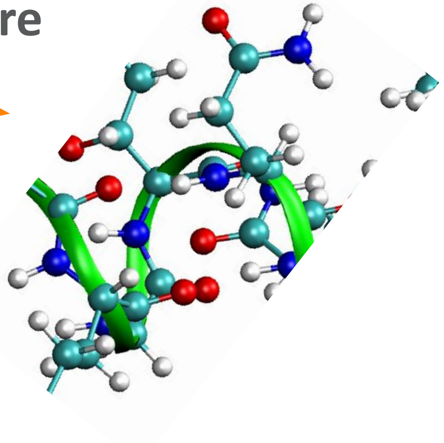
Knowing the 3D structure is critical for drug development

Cell





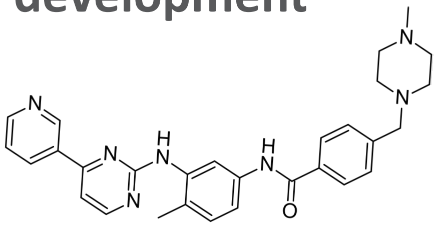
Protein 3D structure



- Information on atomic positions
- Hundreds of thousands of atoms and more

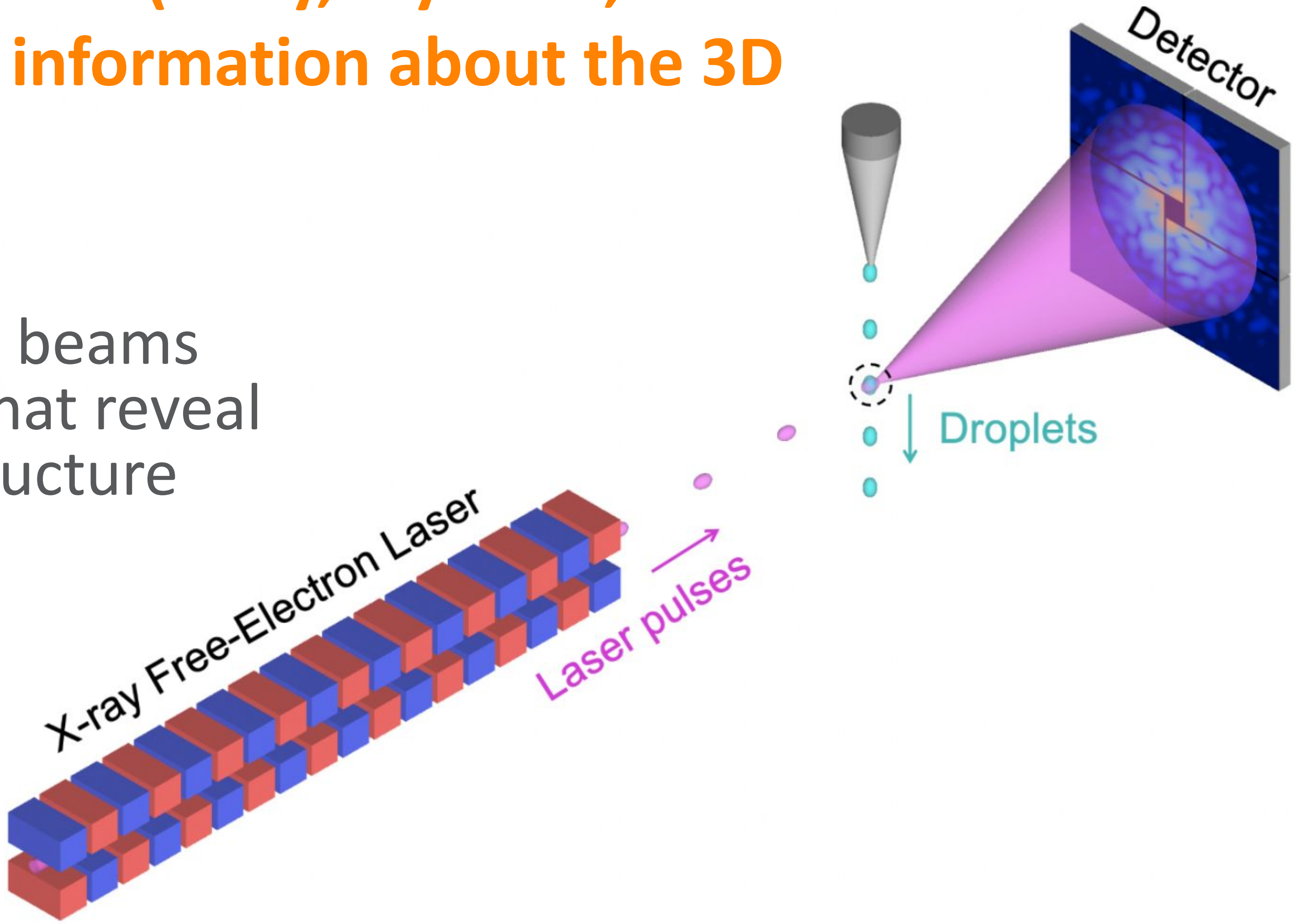
Gleevec® Anticancer drug (Leukemia)

Drug development



There are experimental methods (X-ray, cryo-EM, SAXS, XFEL) to obtain partial information about the 3D protein structure

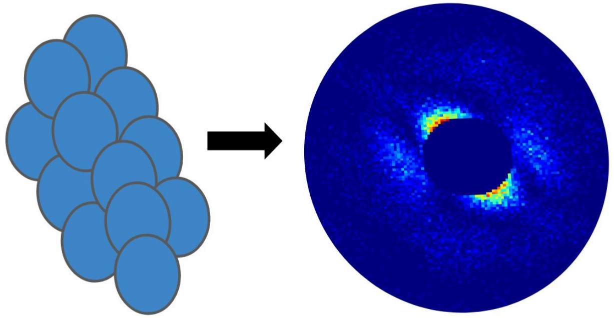
X-ray Free Electron Laser (XFEL) beams create 2D diffraction patterns that reveal properties of the 3D protein structure



Structural properties: Orientation

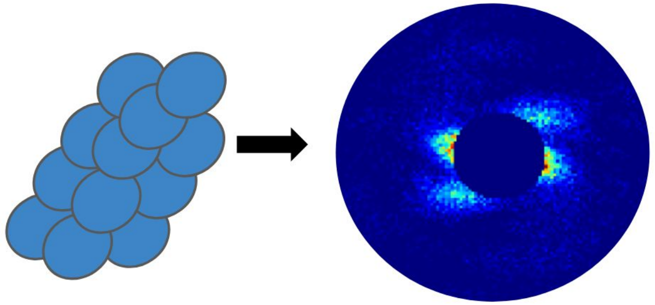
Orientation 1

Φ, θ, Ψ = 24°, 151°, 346°



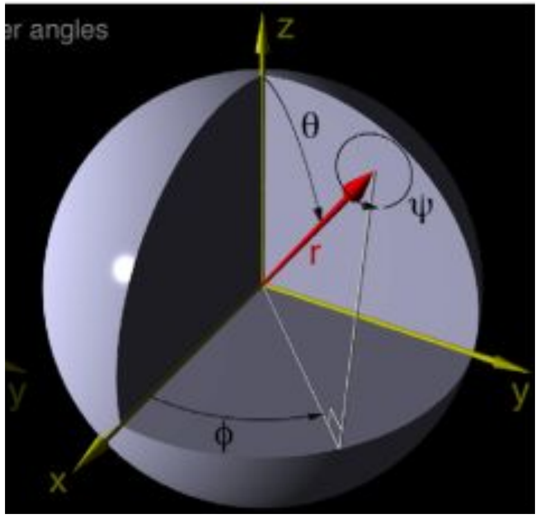
Orientation 2

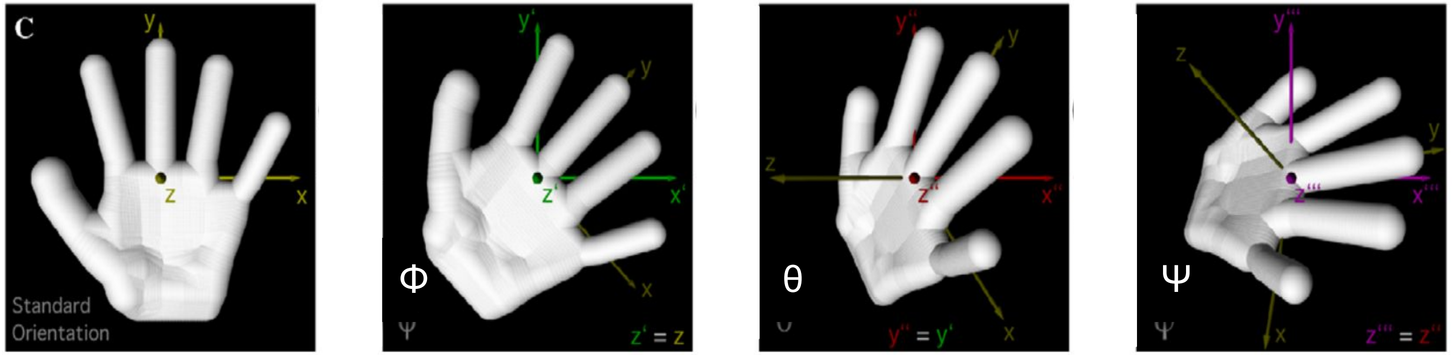
Φ, θ, Ψ = 145°, 128°, 291°



Orientation refers to the **placement of the incident beam** with respect **to a protein structure**

- Φ (Azimuth) = [-180,180]
- Θ (Altitude) = [0,180]
- Ψ (Psi or rotation angle) = [0, 360]





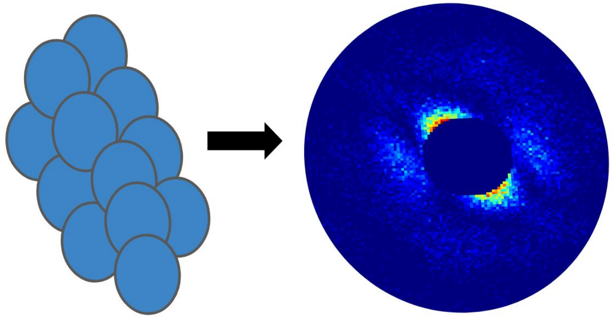
**Images from "Common conventions for interchange and archiving of three-dimensional electron microscopy information in structural biology" by Bernard Heymann, Monica Chagoyen, and David M. Belnap.*

Structural properties: Conformation

Conformation is the **shape adopted by a protein** and is caused by the rotation of the protein atoms around one or more single bonds

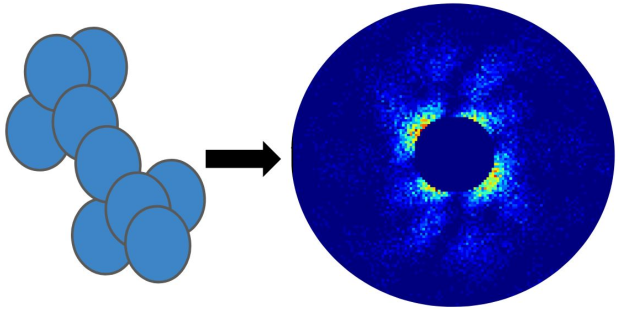
Conformation A

Φ, θ, Ψ = 24°, 151°, 346°



Conformation B

Φ, θ, Ψ = 34°, 139°, 106°



Structural properties: Protein type

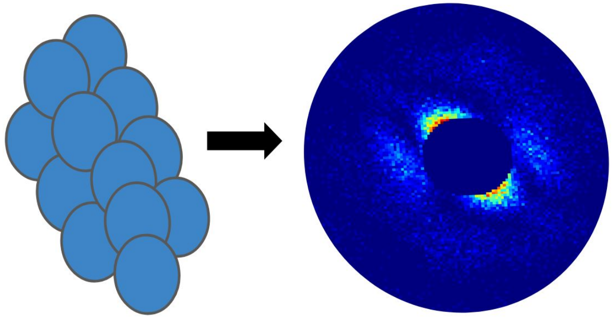
Protein type refers to the type and number of amino acids composing a protein

- 20 different type of amino acids
- Amino acids can combined in different ways to make a protein
  - sequence
  - number (up to thousands)

Protein type A

Conformation A

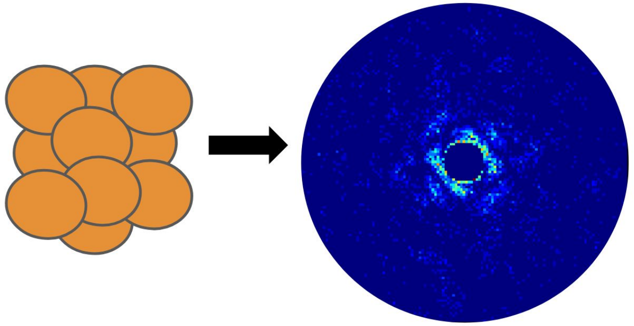
Φ, θ, Ψ = 24°, 151°, 346°



Protein type B

Conformation C

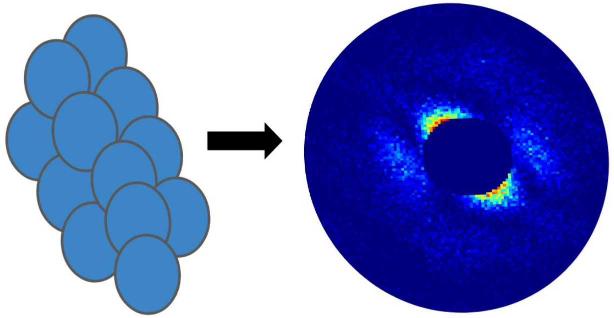
Φ, θ, Ψ = 84°, 32°, 82°



Structural properties

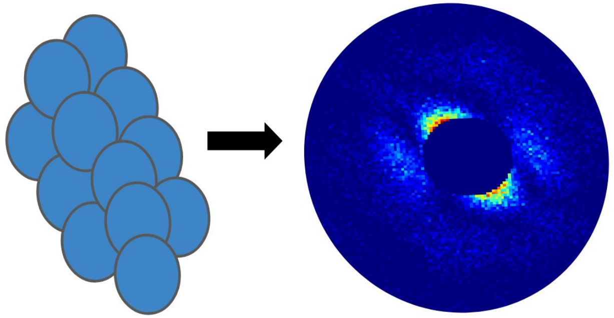
Orientation 1

$\Phi, \theta, \Psi = 24^\circ, 151^\circ, 346^\circ$



Conformation A

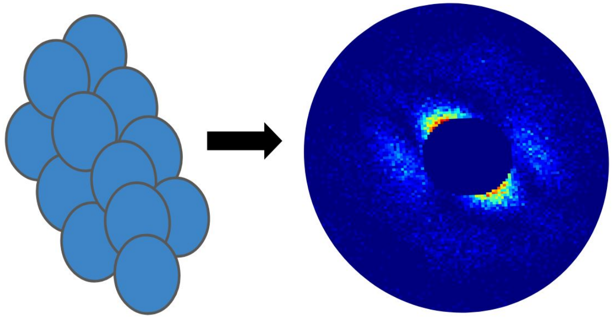
$\Phi, \theta, \Psi = 24^\circ, 151^\circ, 346^\circ$



Protein type A

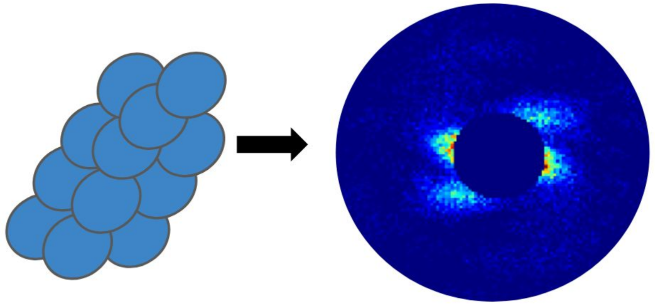
Conformation A

$\Phi, \theta, \Psi = 24^\circ, 151^\circ, 346^\circ$



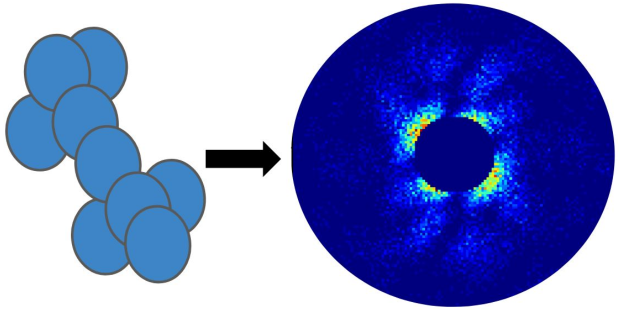
Orientation 2

$\Phi, \theta, \Psi = 145^\circ, 128^\circ, 291^\circ$



Conformation B

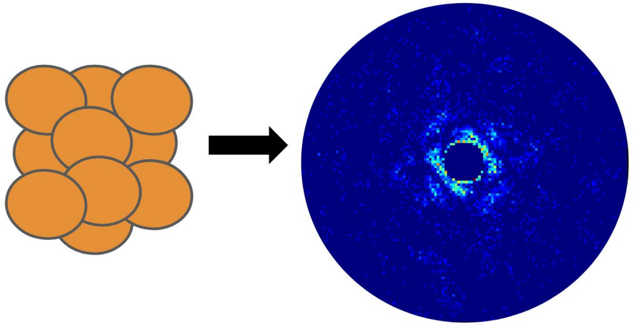
$\Phi, \theta, \Psi = 34^\circ, 139^\circ, 106^\circ$

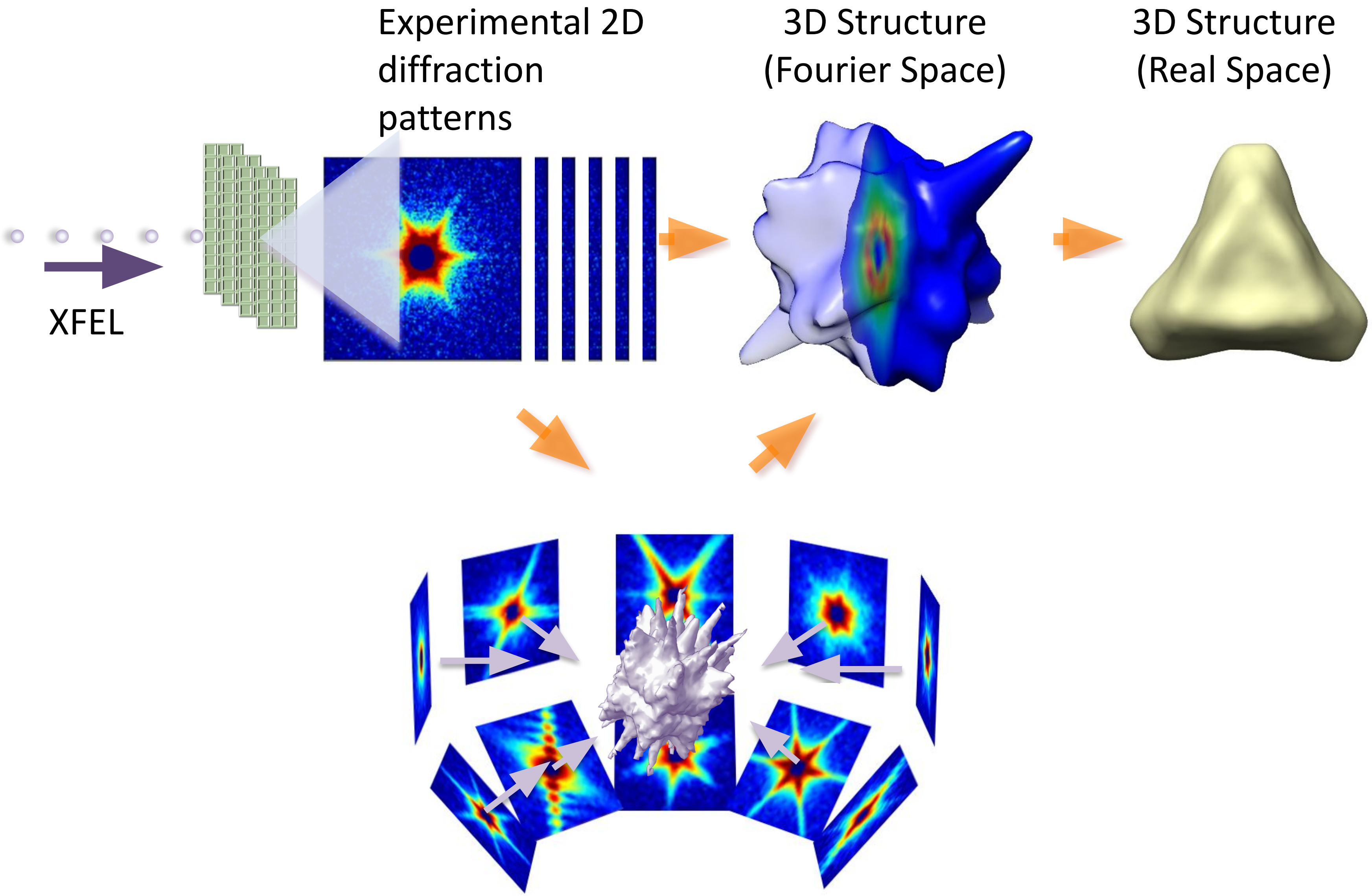


Protein type B

Conformation C

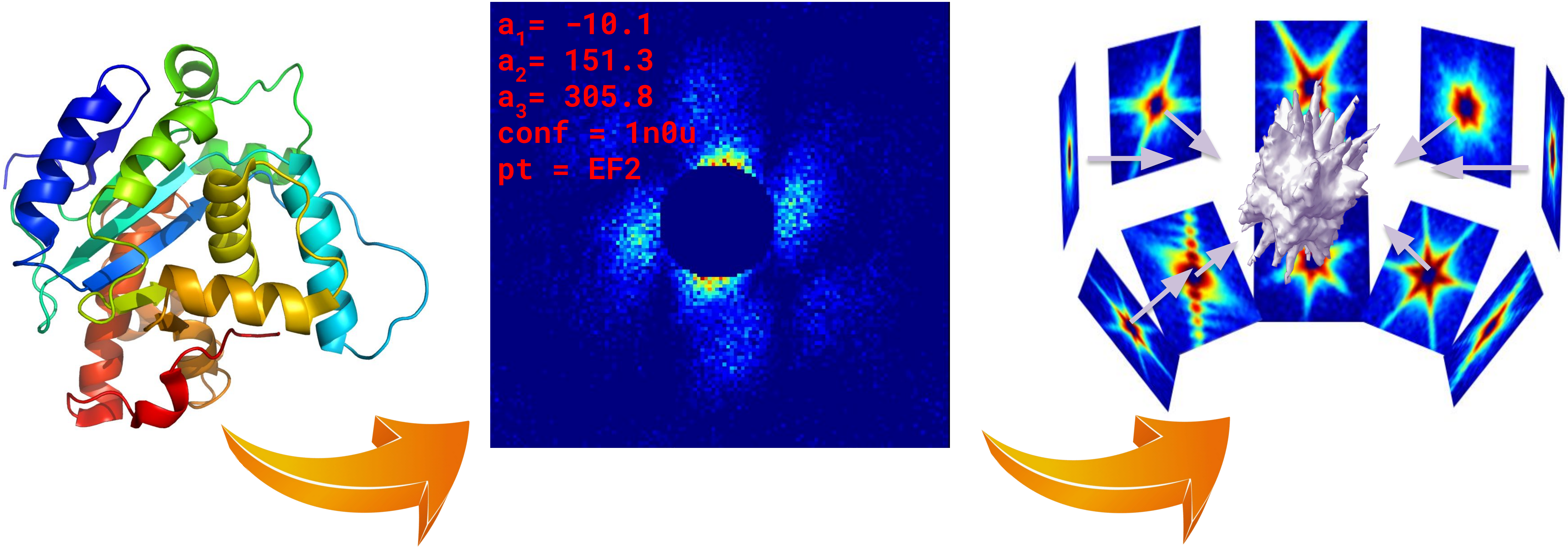
$\Phi, \theta, \Psi = 84^\circ, 32^\circ, 82^\circ$




Experimental 2D diffraction patterns
3D Structure (Fourier Space)
3D Structure (Real Space)
XFEL

Identifying the structural properties embedded in the **2D diffraction pattern** is key for the 3D reconstruction and understanding the protein’s structure

We need to **integrate** the **experimental** methods with **computational** frameworks **to gain information on structure and dynamics** and accelerate scientific discovery



Our goal is to design and implement a **ML-based framework** that predicts simultaneously the three structural properties from protein diffraction patterns

Framework design consideration 1

1. Simultaneous multi-output and multi-type predictions

Orientation

Continuous values

Angle 1 = [-180,180]

Angle 2 = [0,180]

Angle 3 = [0,360]

Conformation

Categorical values

Conf. A1 or Conf. A2 or … or Conformation NN

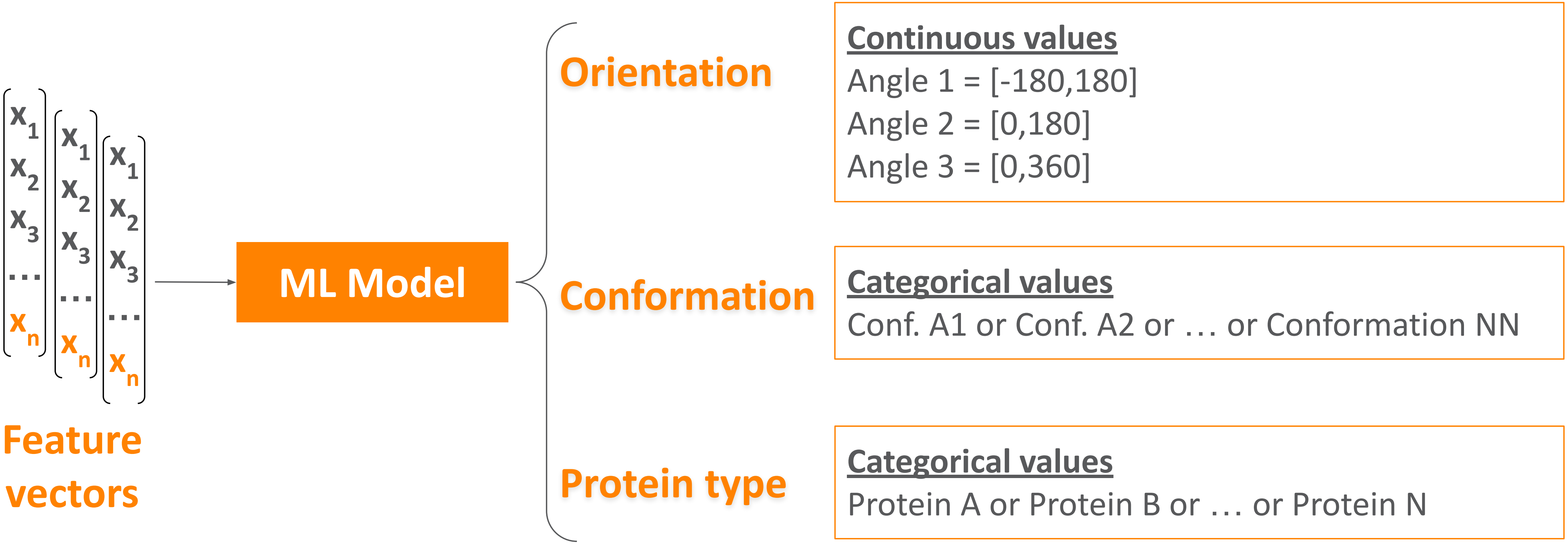
Protein type

Categorical values

Protein A or Protein B or … or Protein N

Framework design consideration 1

1. Simultaneous multi-output and multi-type predictions



Feature vectors

ML Model

Orientation

Continuous values
Angle 1 = [-180,180]
Angle 2 = [0,180]
Angle 3 = [0,360]

Conformation

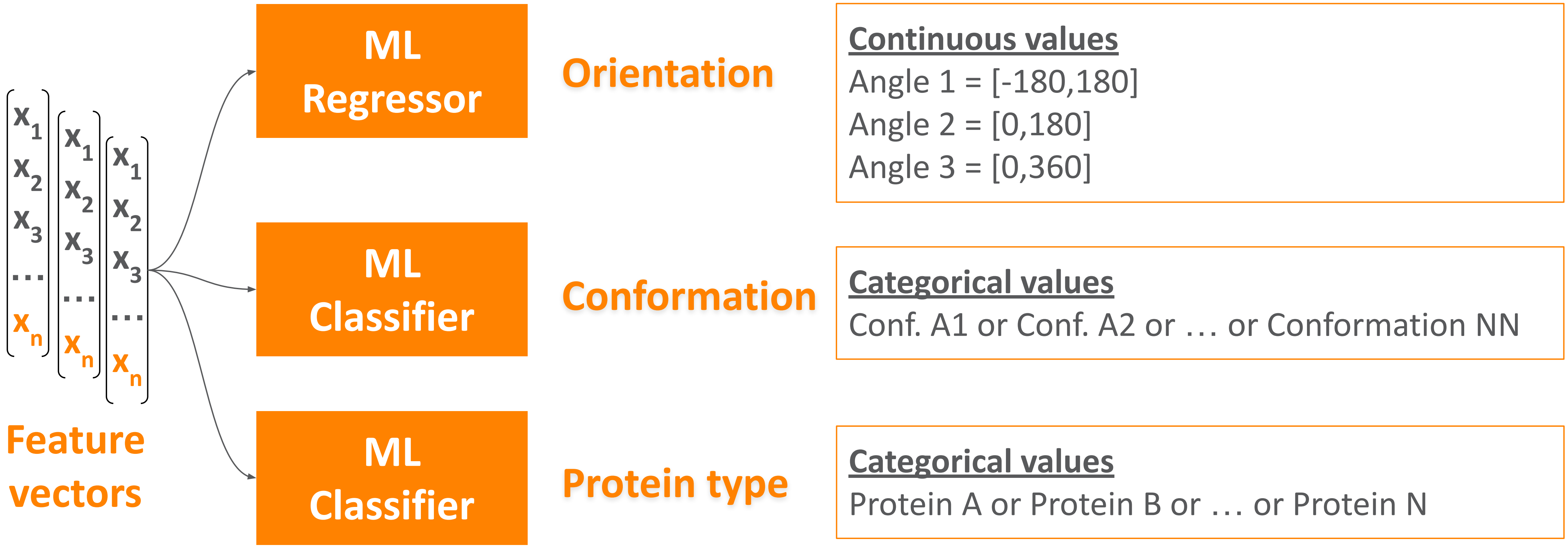
Categorical values
Conf. A1 or Conf. A2 or … or Conformation NN

Protein type

Categorical values
Protein A or Protein B or … or Protein N

Multiple ML models

We define three different ML models for the three predictions



Continuous values

Angle 1 = [-180,180]

Angle 2 = [0,180]

Angle 3 = [0,360]

Categorical values

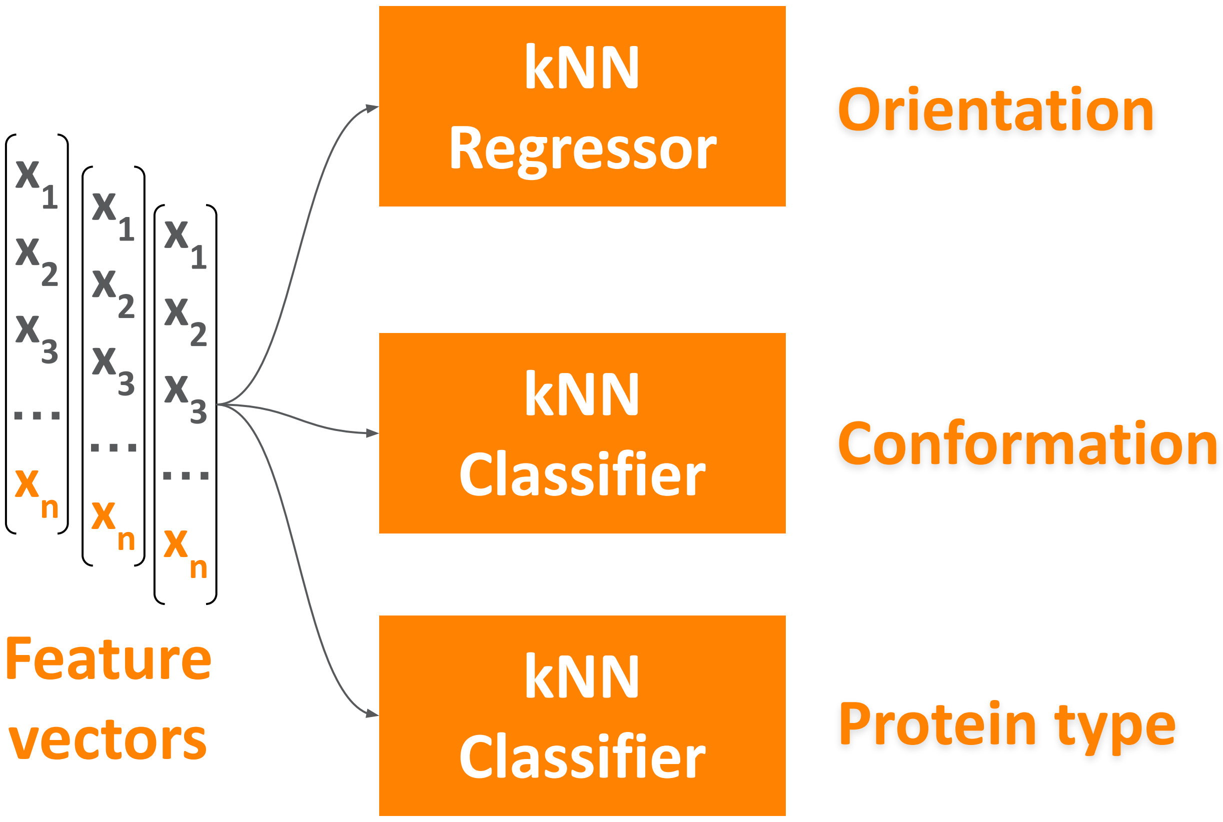
Conf. A1 or Conf. A2 or … or Conformation NN

Categorical values

Protein A or Protein B or … or Protein N

Multiple kNN models

We select kNN (k-Nearest Neighbors) because of it high accuracy in both classification and regression problems and low execution costs



$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix}$$

Feature vectors

kNN Regressor

Orientation

kNN Classifier

Conformation

kNN Classifier

Protein type

Continuous values

Angle 1 = [-180,180]

Angle 2 = [0,180]

Angle 3 = [0,360]

Categorical values

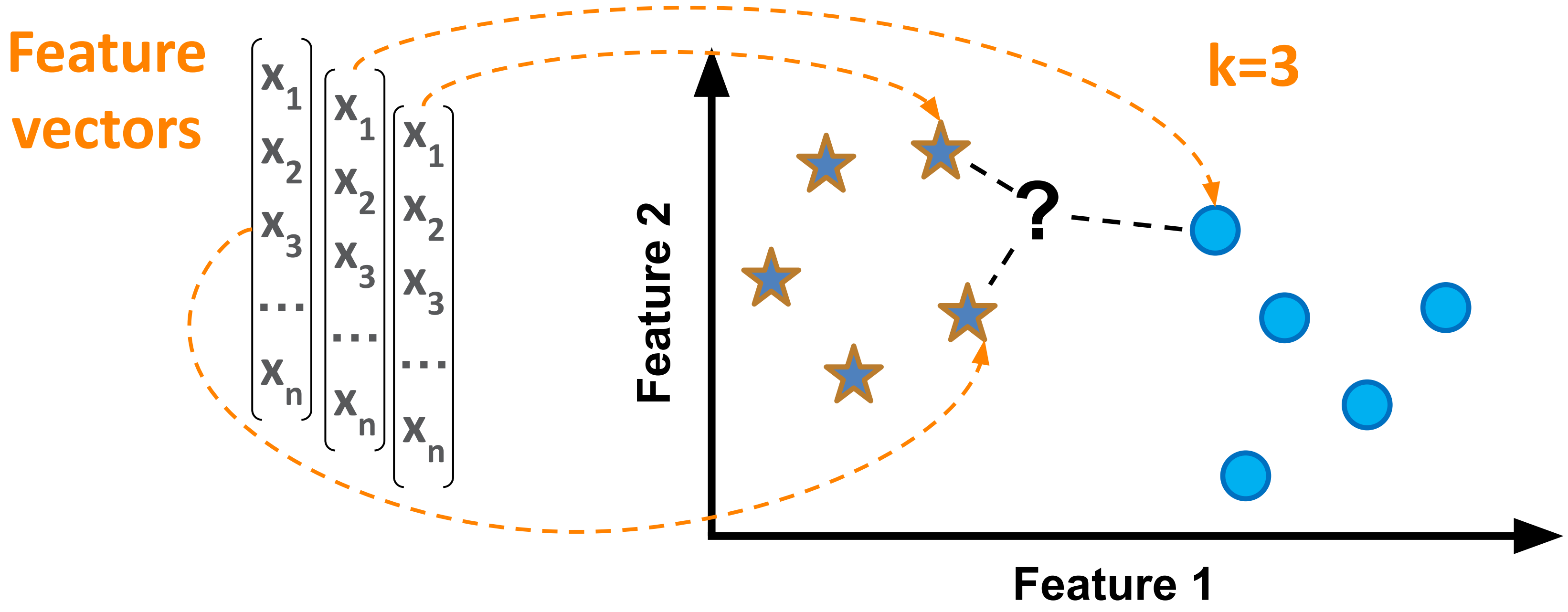
Conf. A1 or Conf. A2 or … or Conformation NN

Categorical values

Protein A or Protein B or … or Protein N

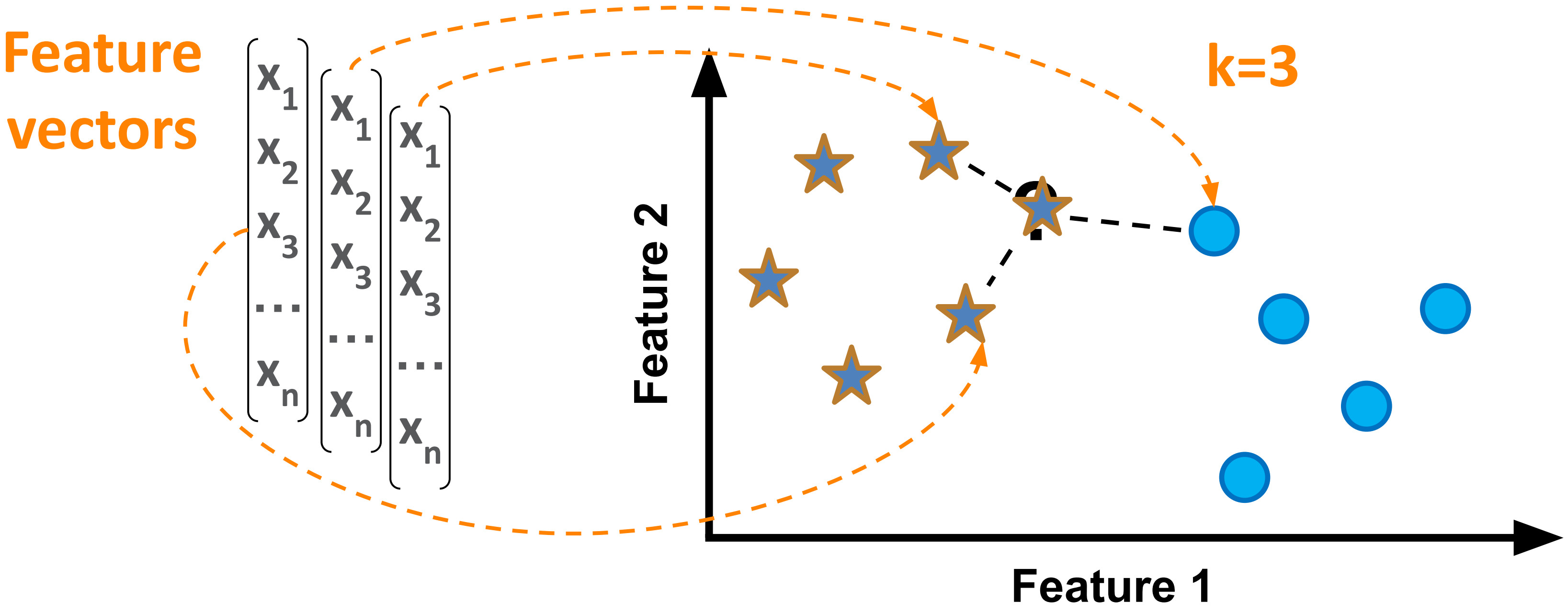
k-Nearest Neighbors (kNN)

This algorithm looks at the *K* nearest neighbors of a new data point (in feature space) to determine the predicted value



k-Nearest Neighbors (kNN)

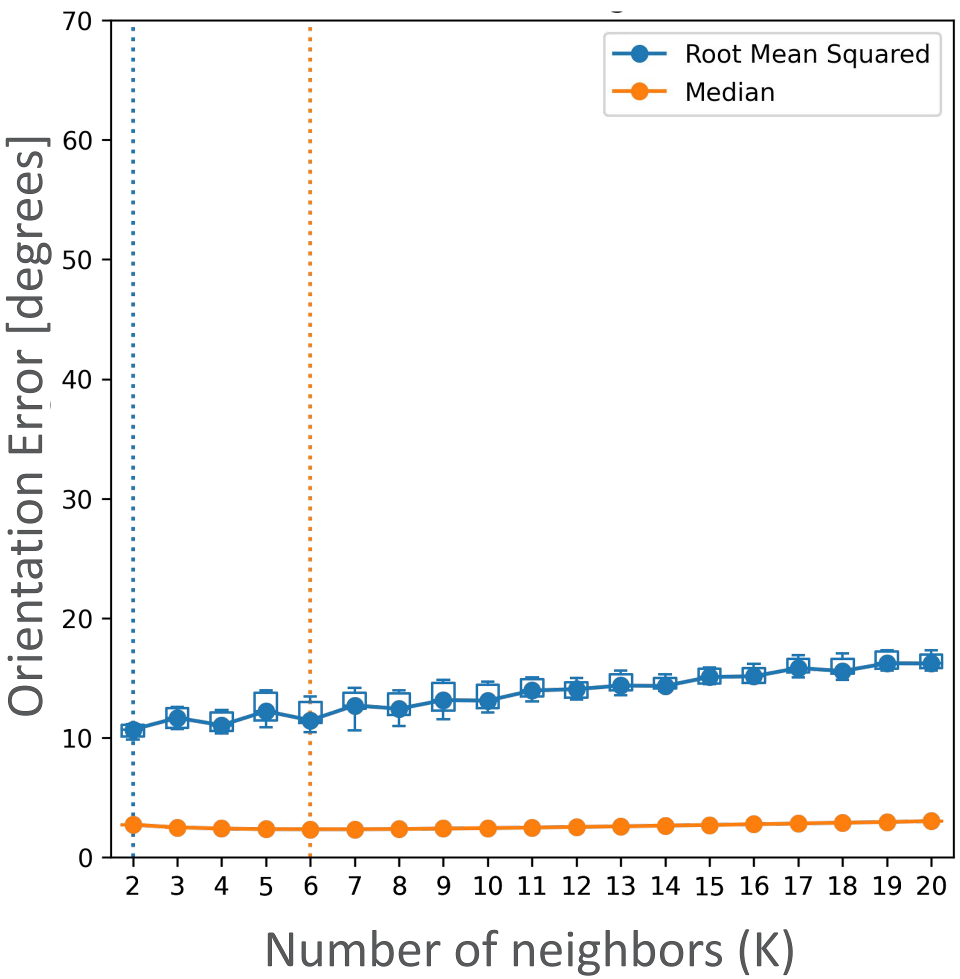
This algorithm looks at the *K* nearest neighbors of a new data point (in feature space) to determine the predicted value


Feature vectors
x_1 x_2 x_3 … x_n
x_1 x_2 x_3 … x_n
x_1 x_2 x_3 … x_n
k=3
?
Feature 2
Feature 1

The K number of neighbors is critical for the prediction

Selecting the K number of neighbors

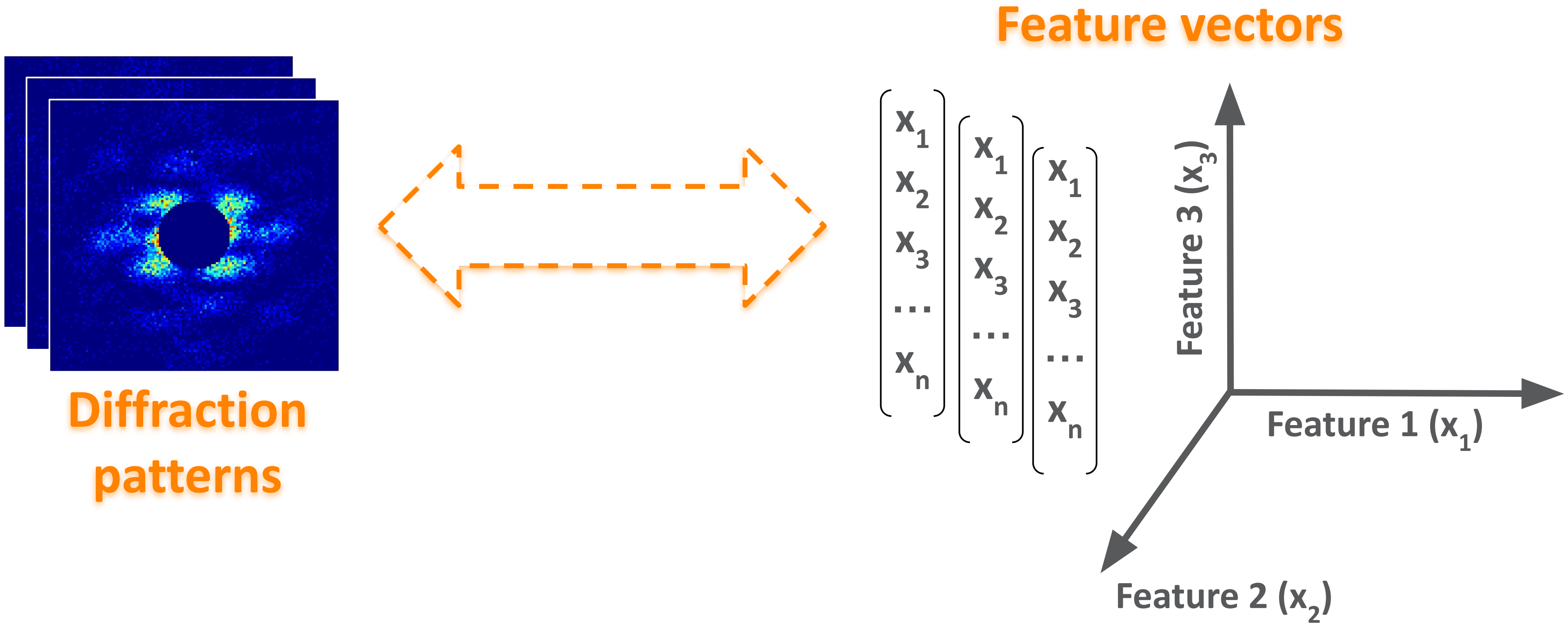
The K number of neighbors is critical for the prediction



An analysis of the root mean square error (RMSE) of the degree allows our framework to identify the most suitable K number of neighbors

Framework design consideration 2

2. Transformation from diffraction patterns to feature vectors

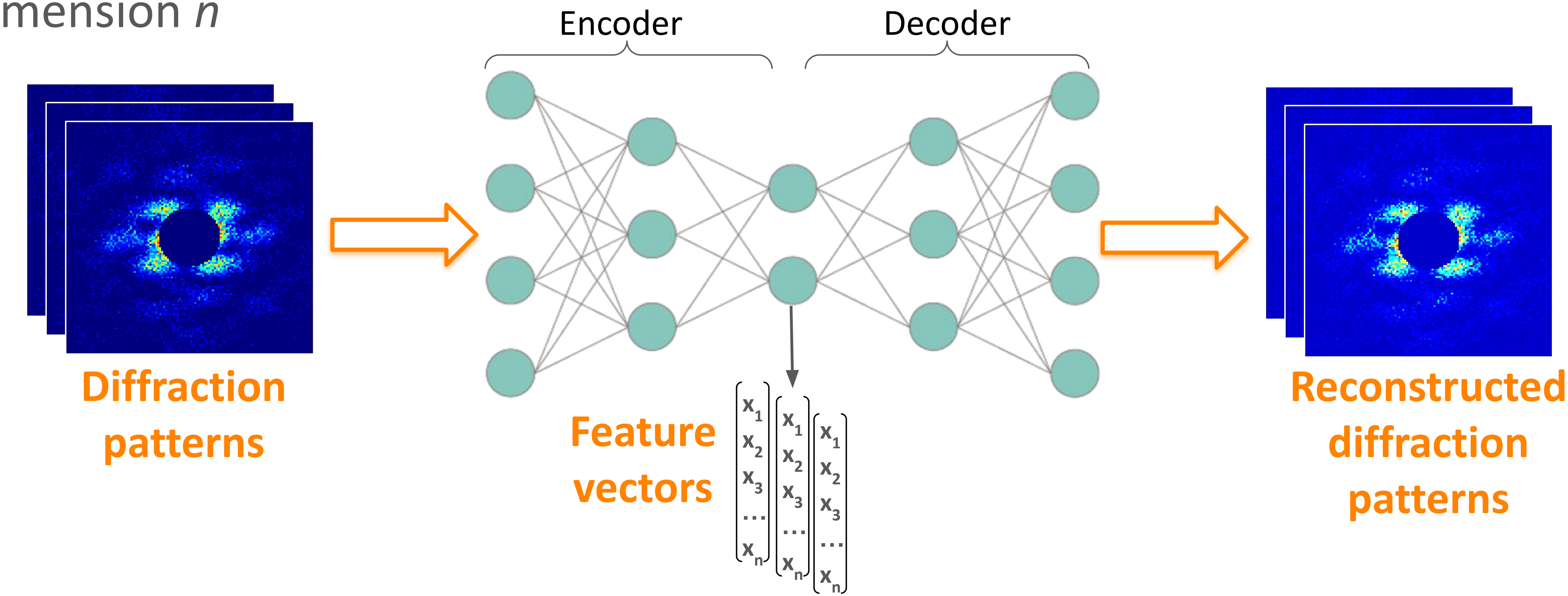


Feature vectors

Diffraction patterns

From diffraction patterns to feature vectors

We use an **autoencoder** to represent diffraction patterns in feature vectors of dimension *n*



Encoder

Decoder

Diffraction patterns

Feature vectors

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \cdots \\ x_n \end{bmatrix}$$

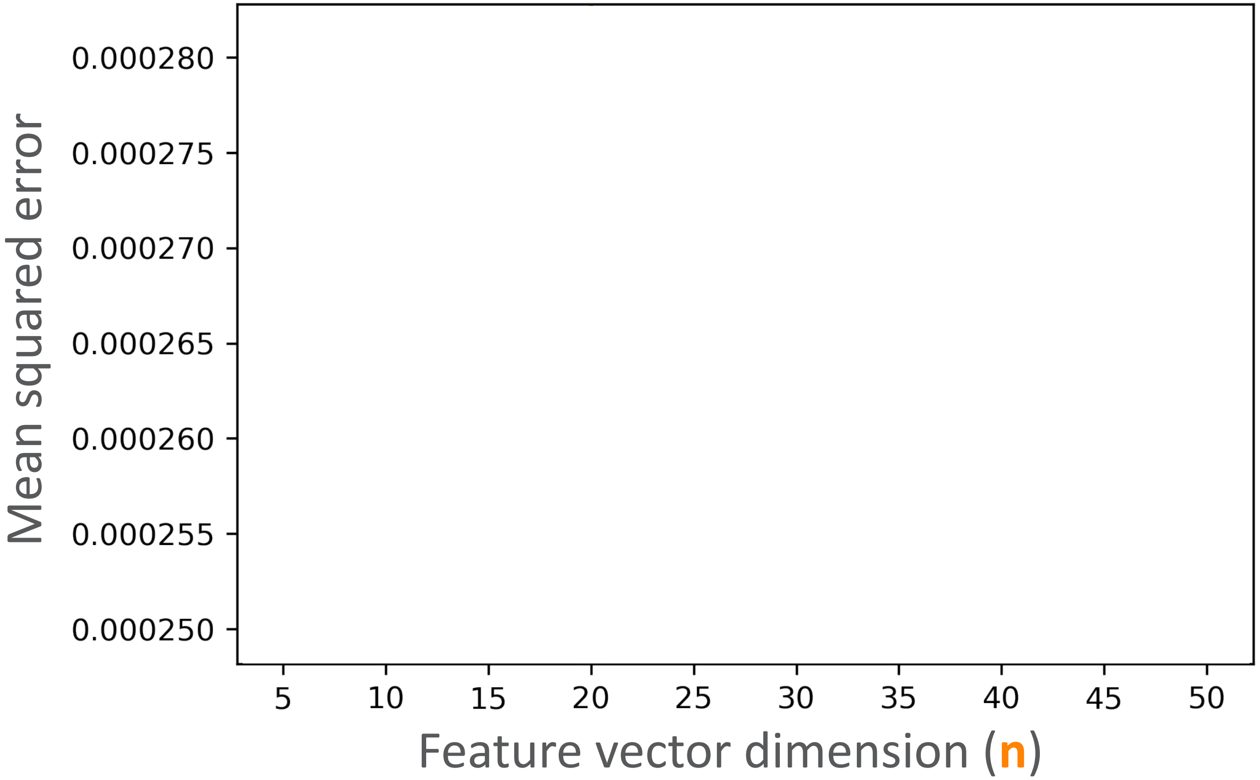
Reconstructed diffraction patterns

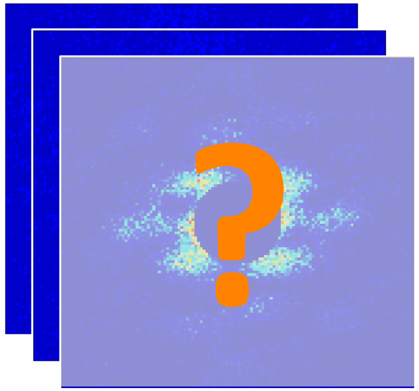
Feature vectors dimension (n)

The dimension of the feature vector has to be sufficient to faithfully reconstruct the original diffraction patterns

Feature vectors

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \dots \\ x_n \end{bmatrix}$$

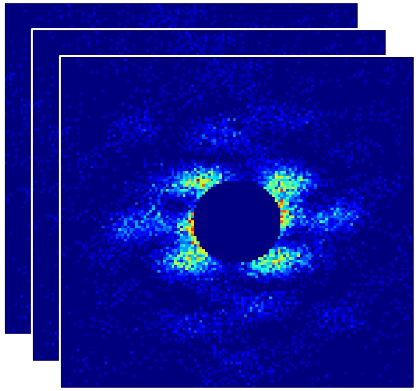




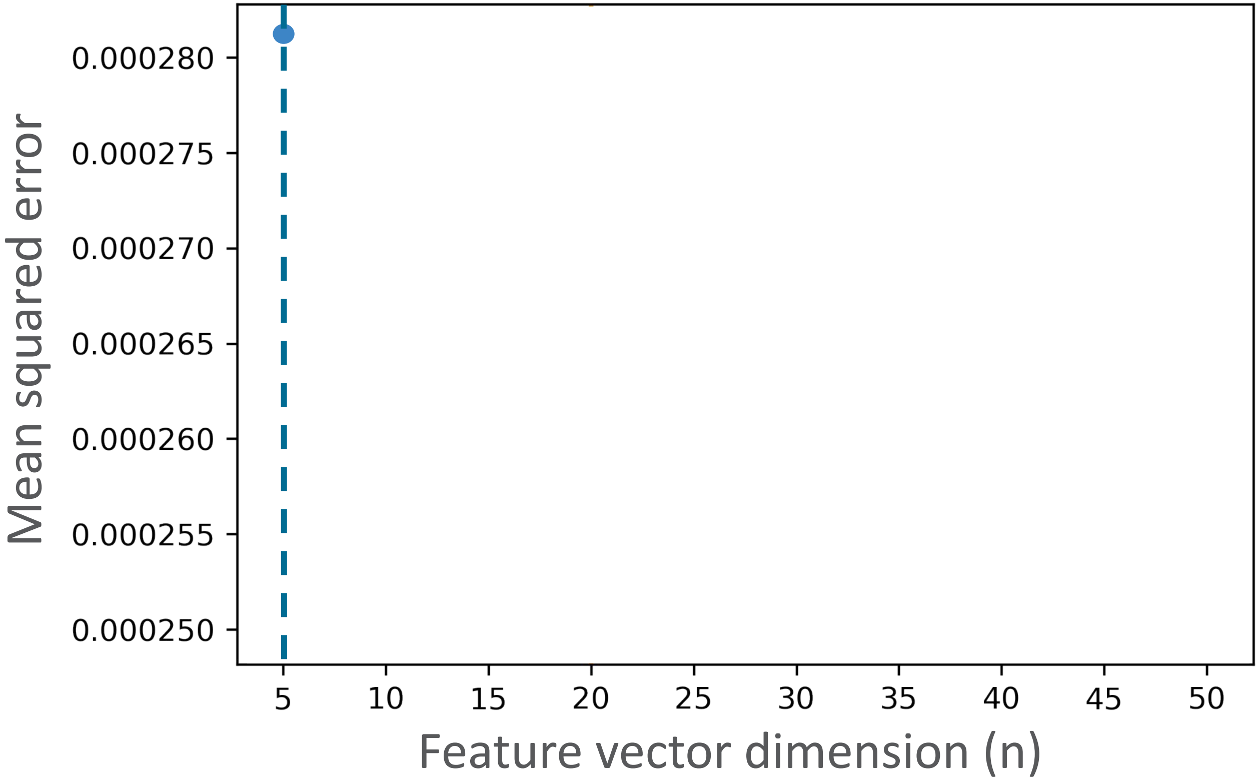
Reconstructed diffraction patterns

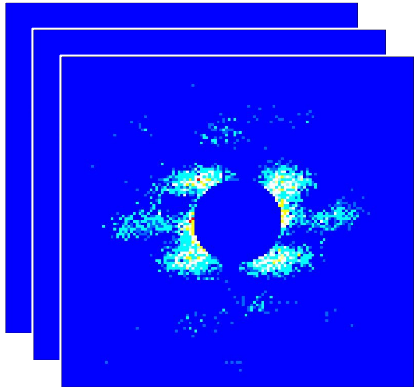
Feature vectors dimension (n=5)

The dimension of the feature vector has to be sufficient to faithfully reconstruct the original diffraction patterns



Diffraction patterns

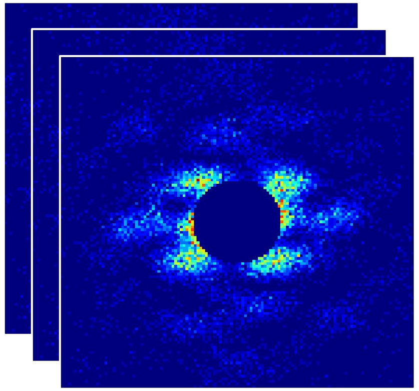




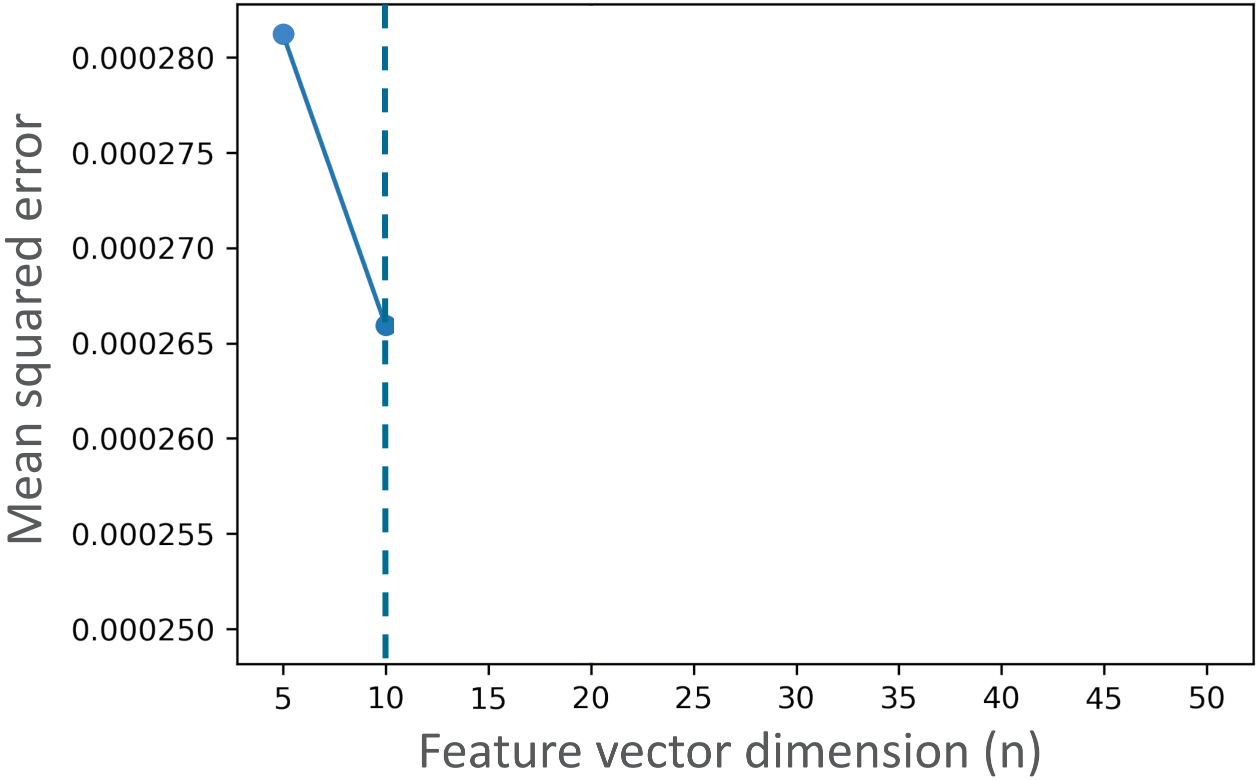
Reconstructed diffraction patterns

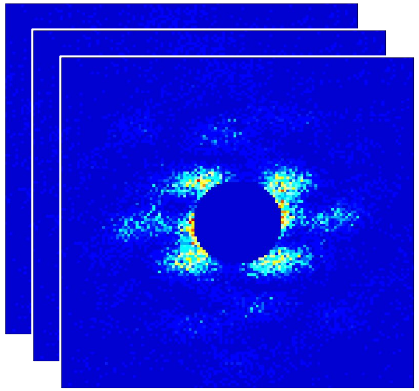
Feature vectors dimension (n=10)

The dimension of the feature vector has to be sufficient to faithfully reconstruct the original diffraction patterns



Diffraction patterns

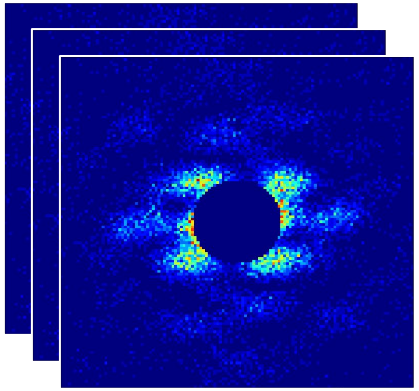




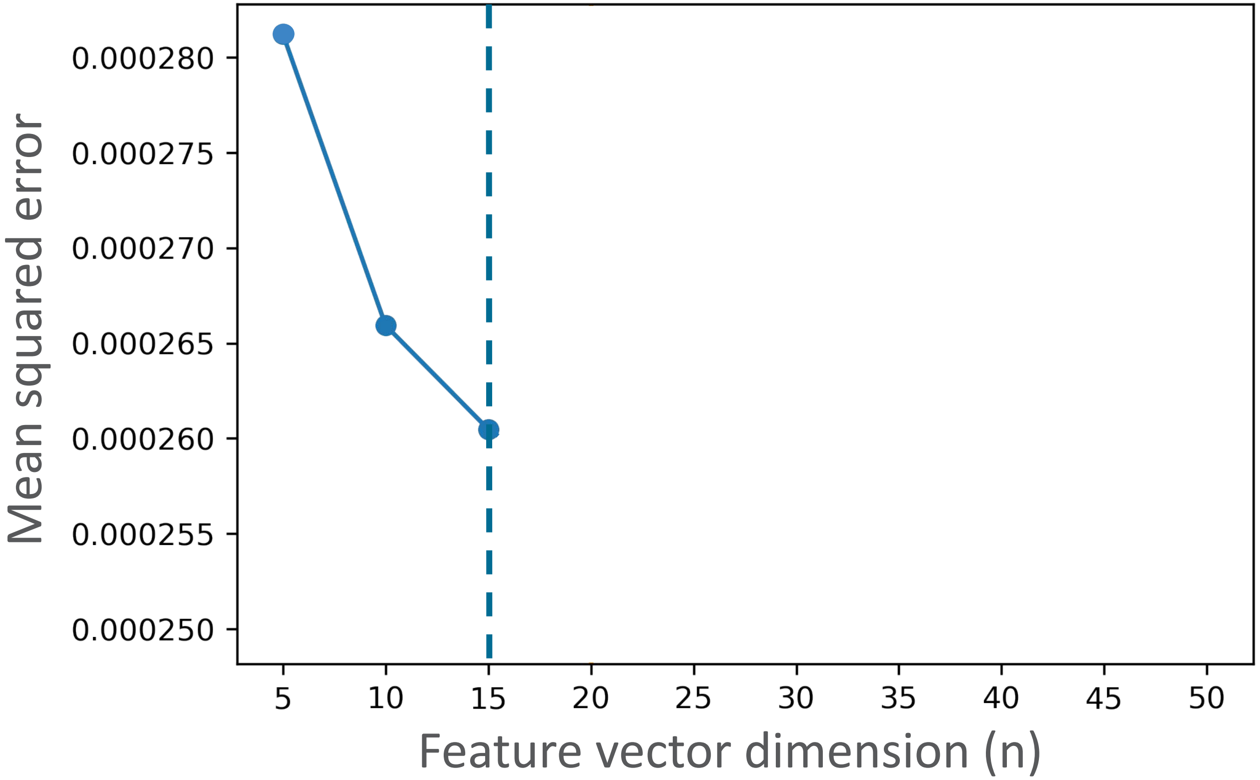
Reconstructed diffraction patterns

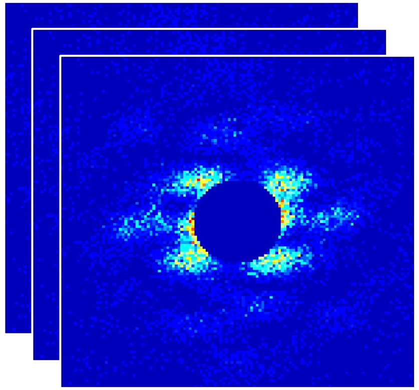
Feature vectors dimension (n=15)

The dimension of the feature vector has to be sufficient to faithfully reconstruct the original diffraction patterns



Diffraction patterns

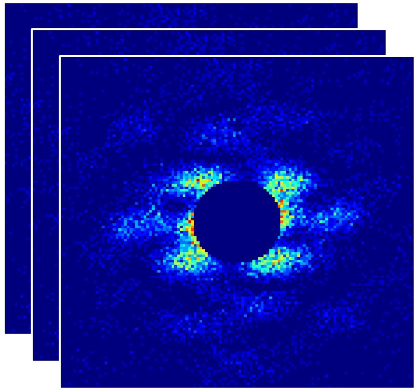




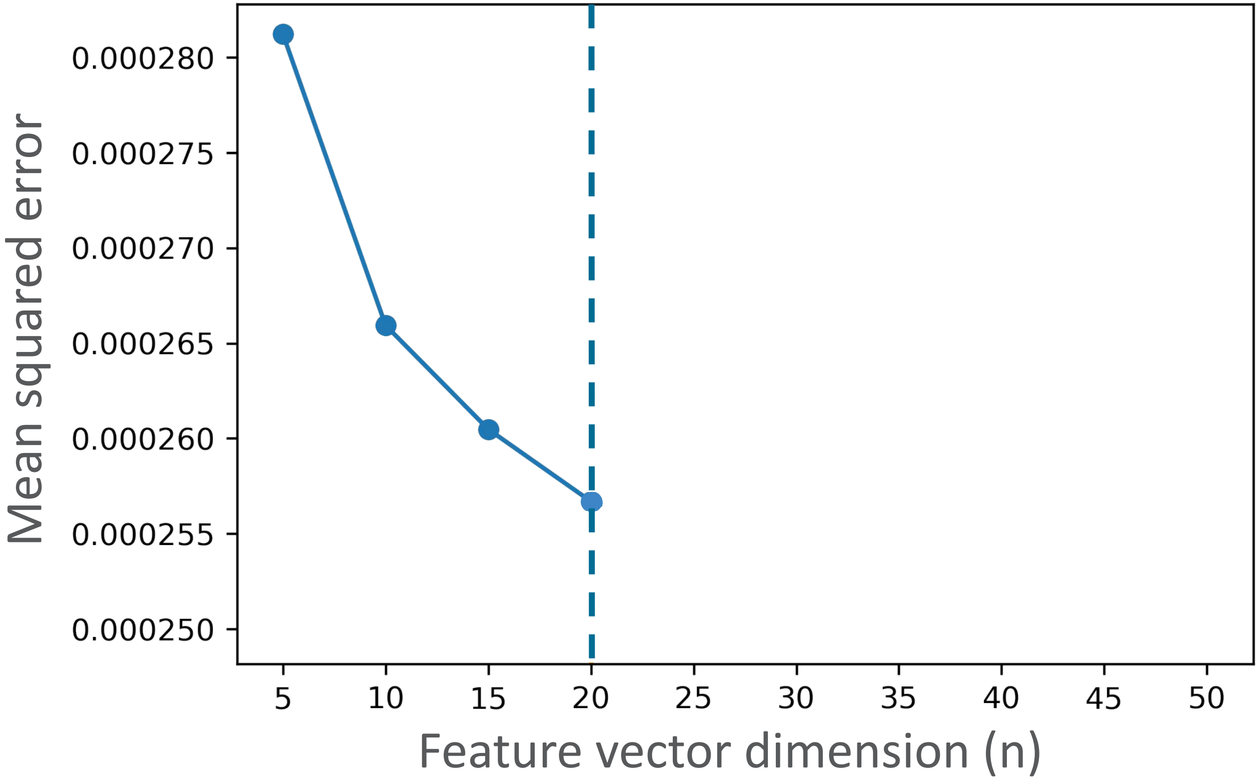
Reconstructed diffraction patterns

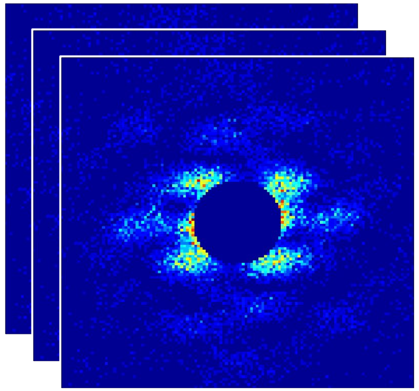
Feature vectors dimension (n=20)

The dimension of the feature vector has to be sufficient to faithfully reconstruct the original diffraction patterns



Diffraction patterns

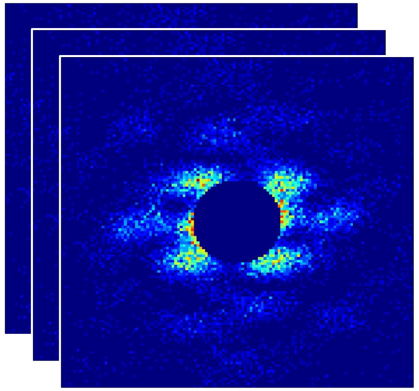




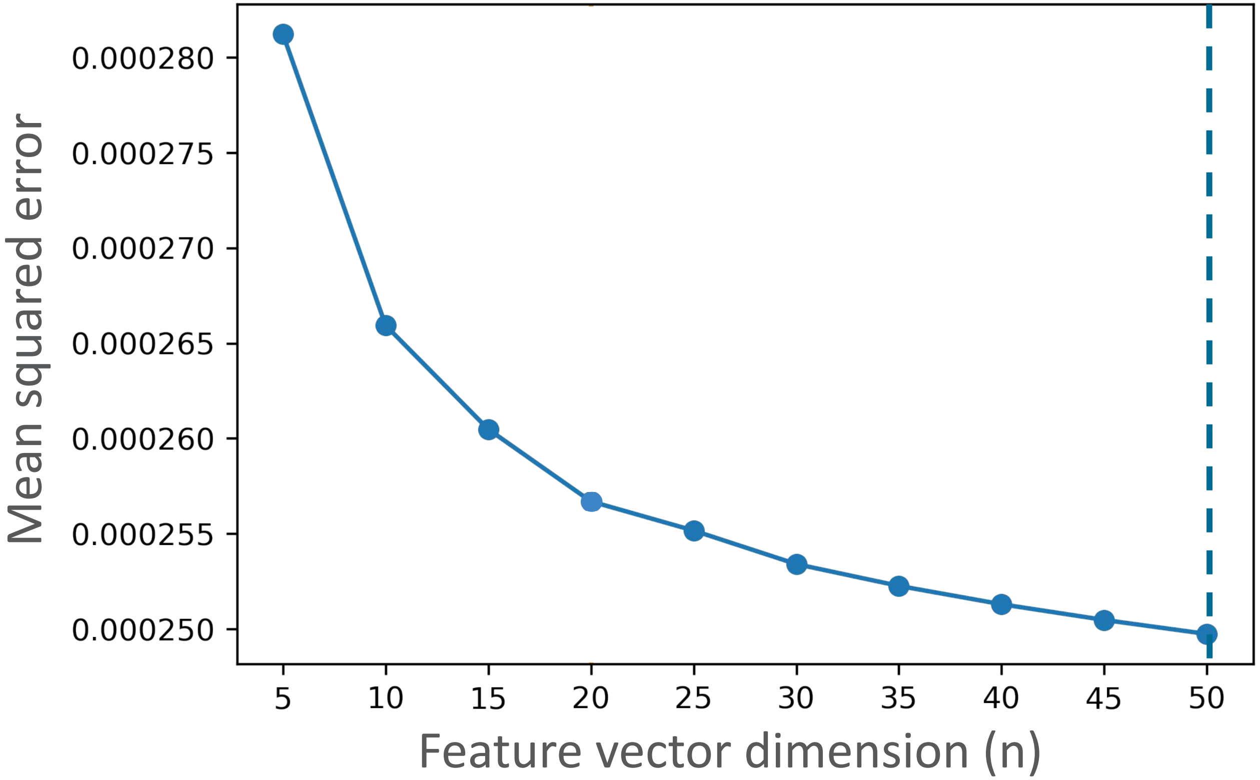
Reconstructed diffraction patterns

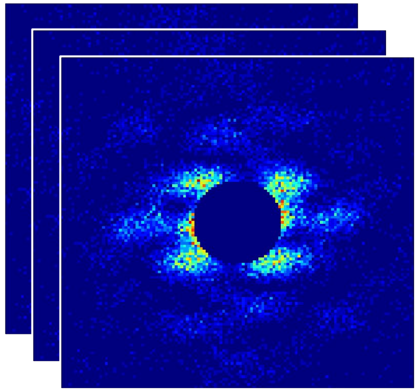
Feature vectors dimension (n=50)

The dimension of the feature vector has to be sufficient to faithfully reconstruct the original diffraction patterns



Diffraction patterns





Reconstructed diffraction patterns

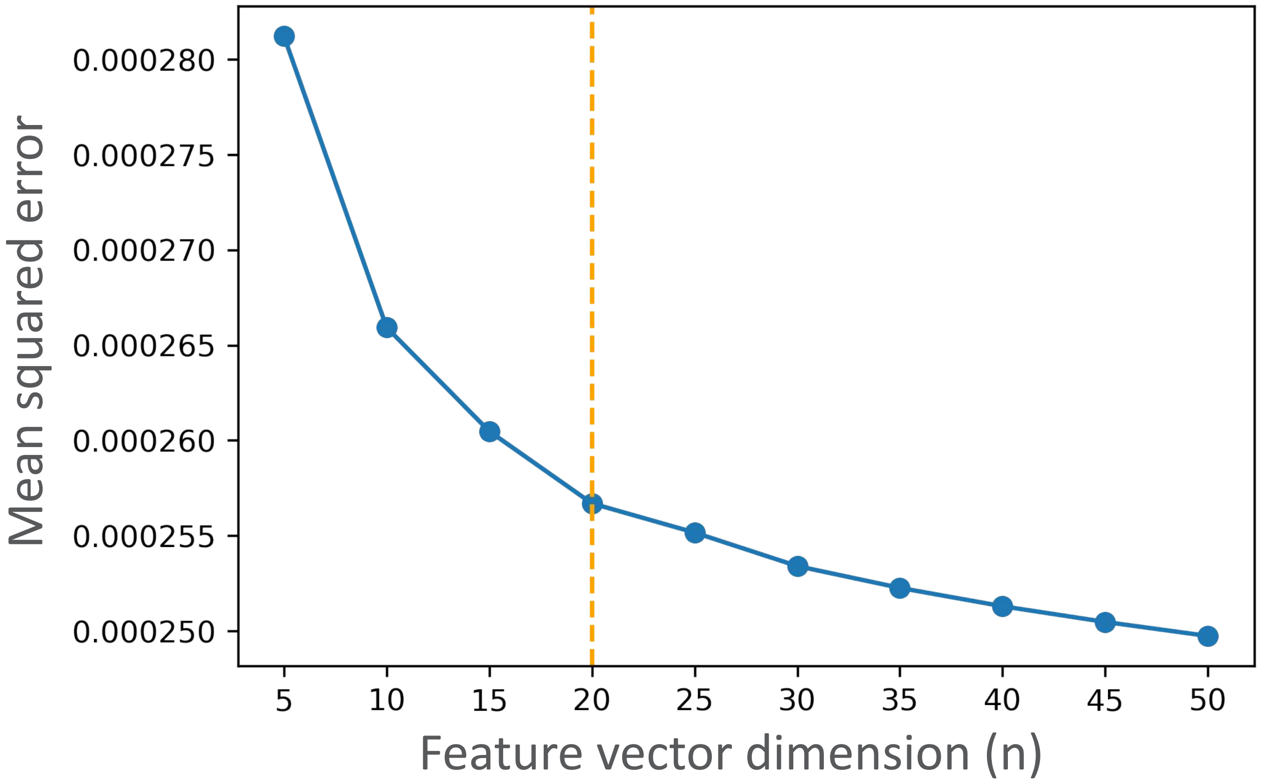
Identify the suitable feature vector dimension

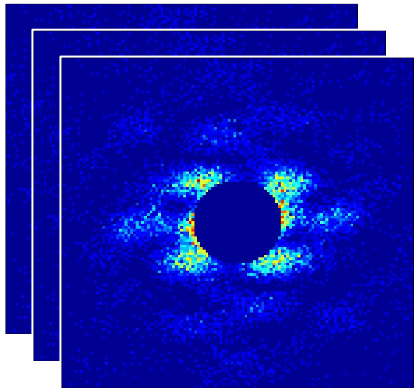
Using the elbow method, our framework identifies when variance of the error and the associated gain in accuracy are not significant

Feature vectors

$$\begin{bmatrix}x_1\\x_2\\x_3\\\cdots\\x_{20}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\\cdots\\x_{20}\end{bmatrix}\begin{bmatrix}x_1\\x_2\\x_3\\\cdots\\x_{20}\end{bmatrix}$$

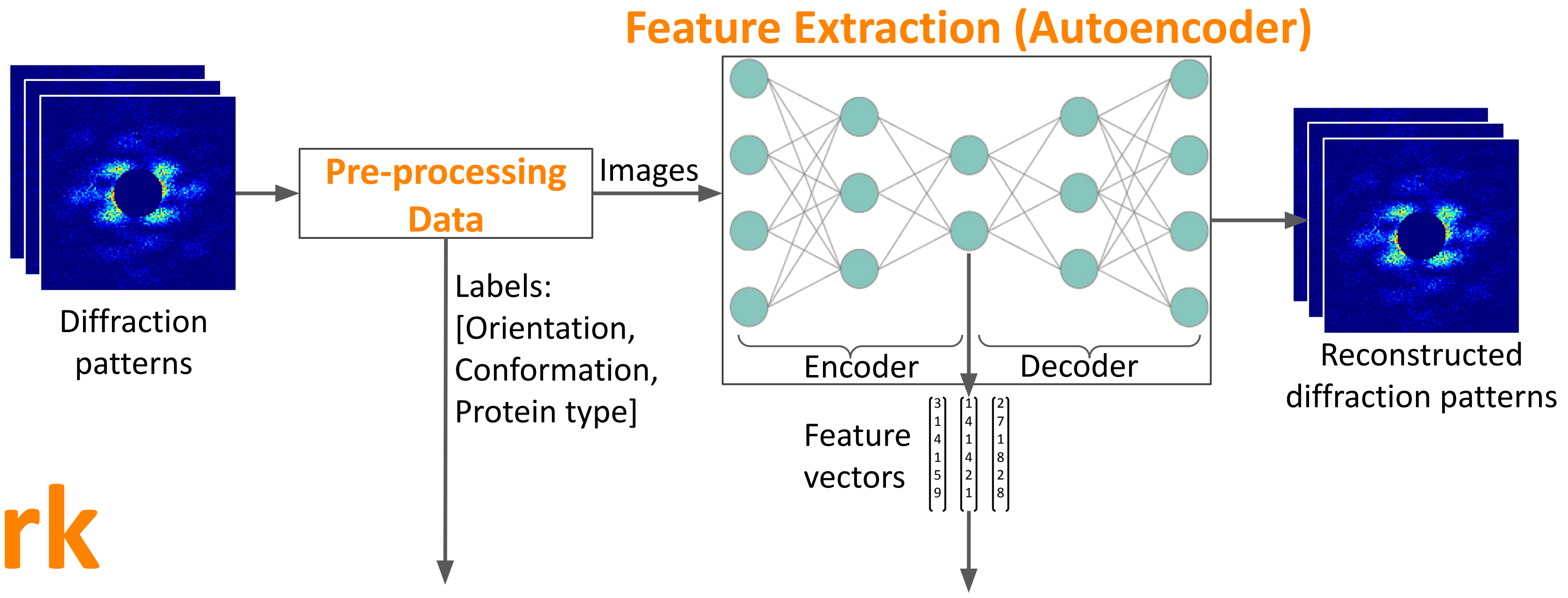
n=20





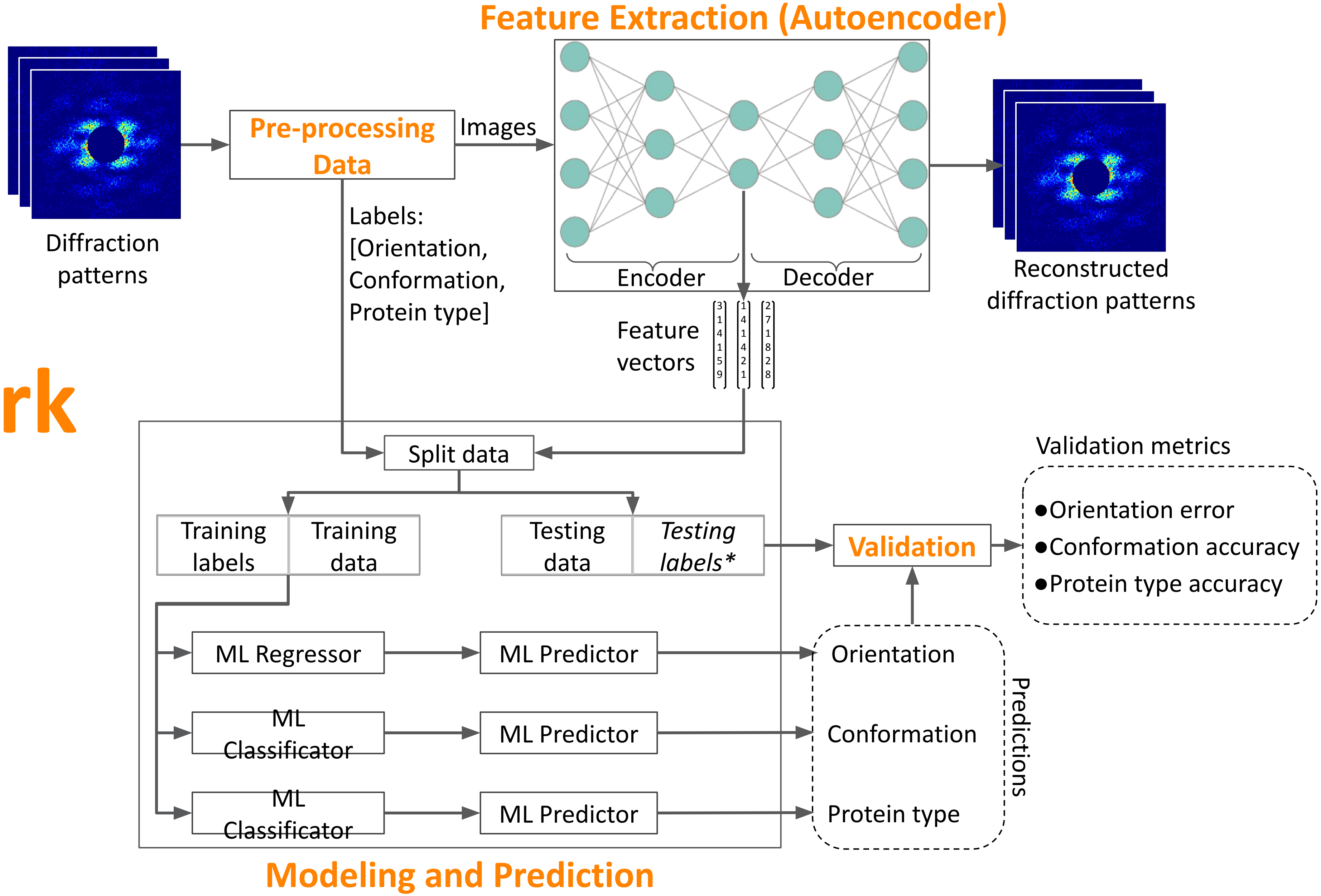
Reconstructed diffraction patterns

XPSI Framework


Feature Extraction (Autoencoder)
Diffraction patterns
Pre-processing Data
Images
Labels: [Orientation, Conformation, Protein type]
Encoder
Decoder
Feature vectors
3 1 4 1 5 9
1 4 1 4 2 1
2 7 1 8 2 8
Reconstructed diffraction patterns

XPSI Framework

XPSI identifies structural properties (i.e., orientation, conformation, protein type)



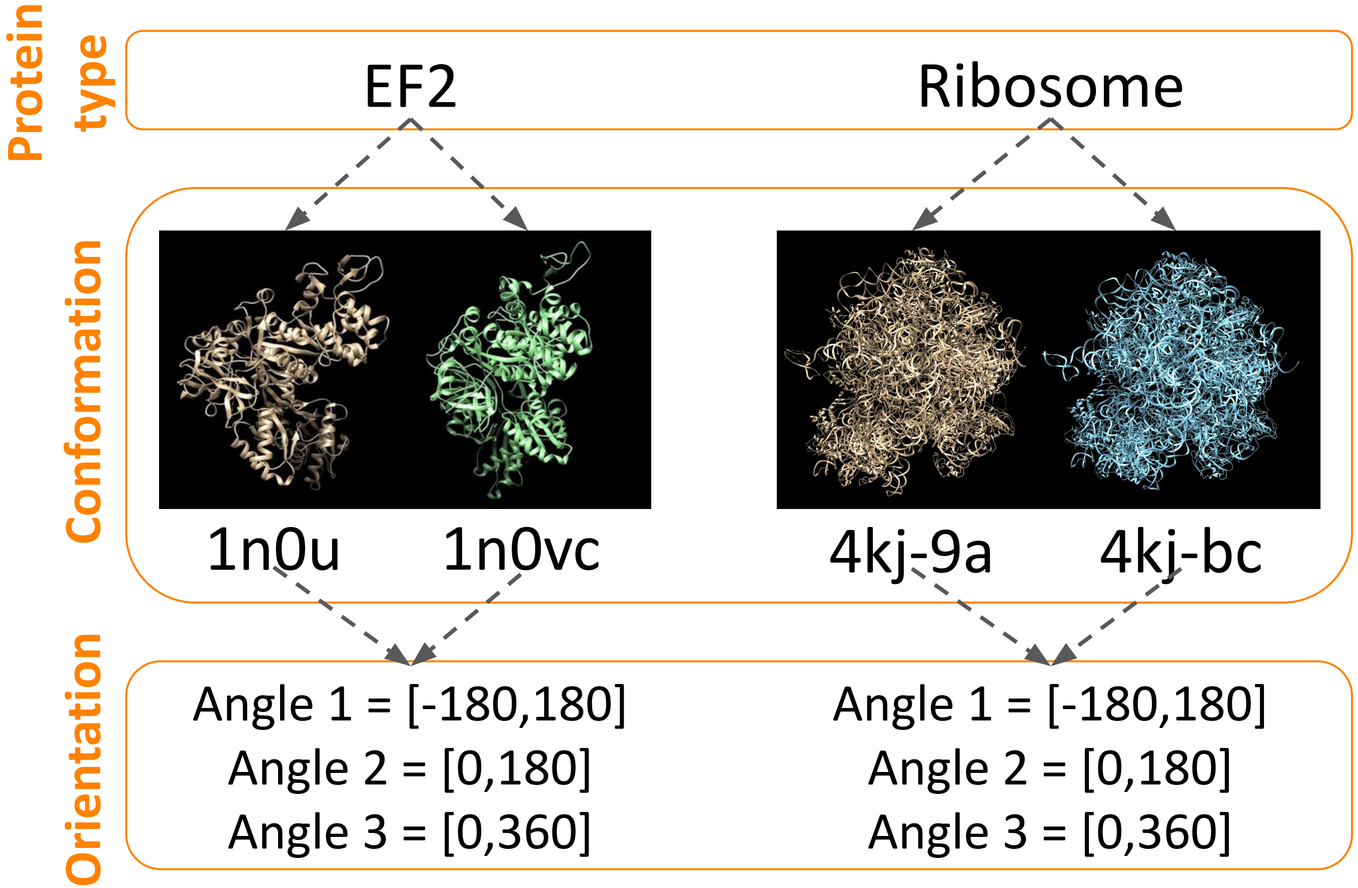
Validation metrics

- Orientation error
- Conformation accuracy
- Protein type accuracy

Identifying structural properties with XPSI

We demonstrate our framework's capability to identify structural properties by merging diverse datasets of diffraction patterns with multiple orientations, conformations, and protein types

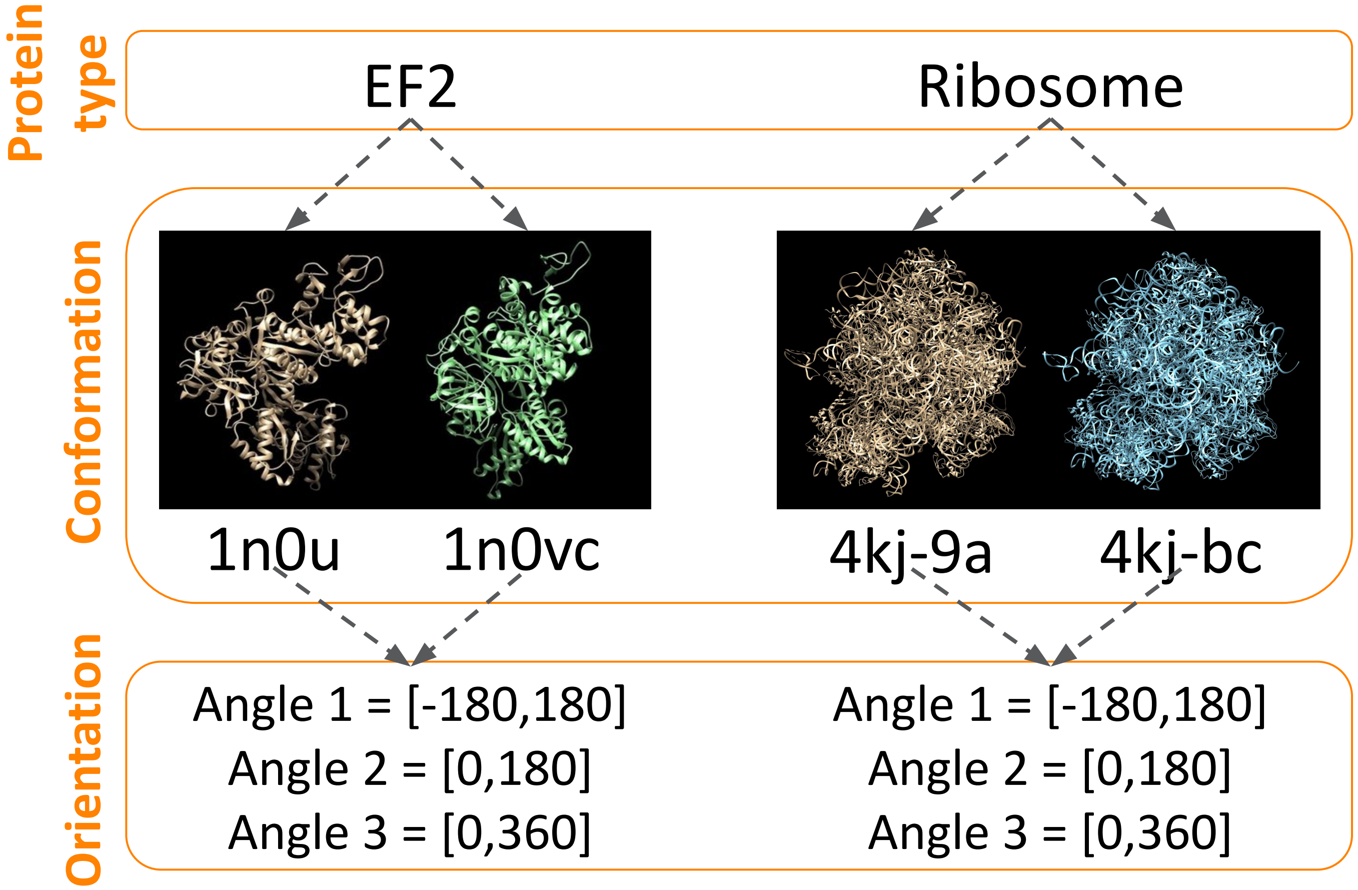
- 39,692 diffraction patterns per each conformation



Identifying structural properties with XPSI

We demonstrate our framework's capability to identify structural properties by merging diverse datasets of diffraction patterns with multiple orientations, conformations, and protein types

But there is one extra challenge … Noise



Protein type

EF2

Ribosome

Conformation

1n0u

1n0vc

4kj-9a

4kj-bc

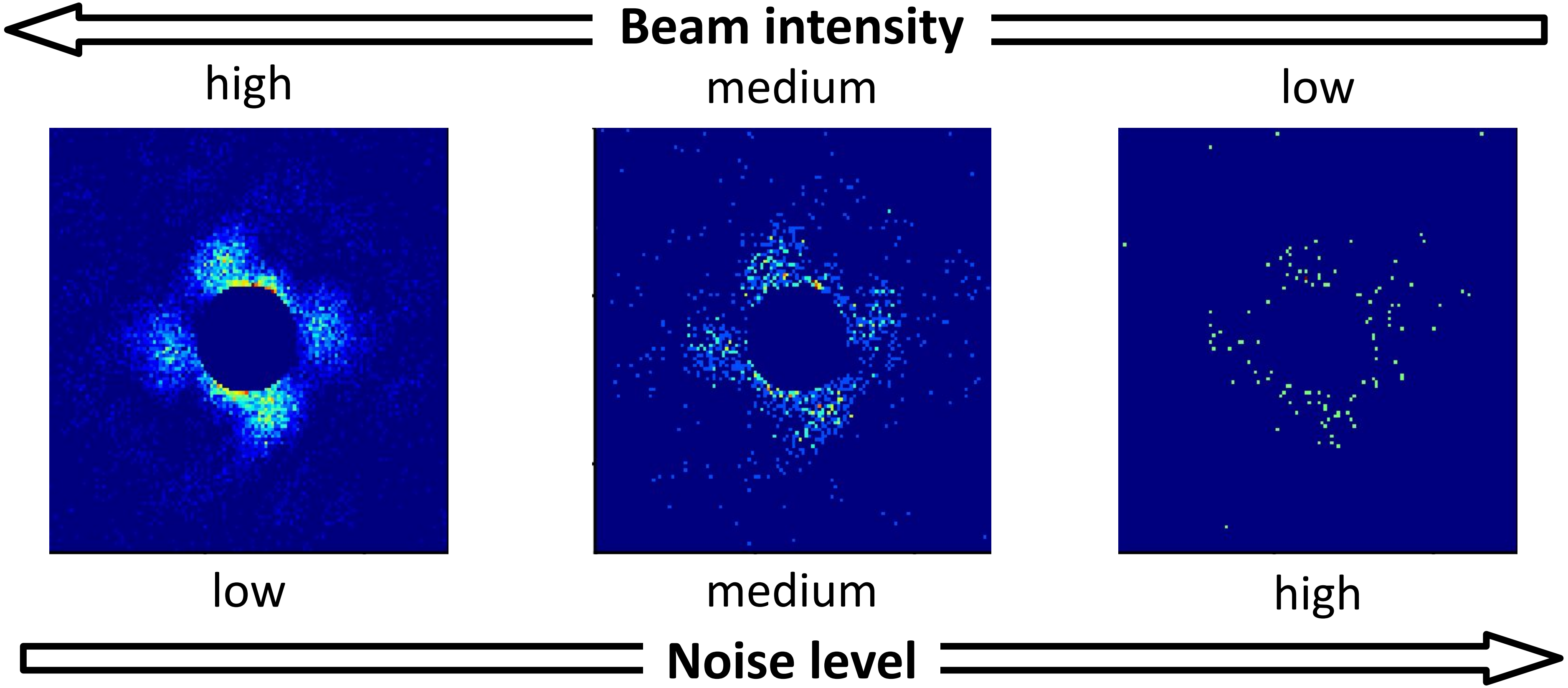
Orientation

Angle 1 = [-180,180]
Angle 2 = [0,180]
Angle 3 = [0,360]

Angle 1 = [-180,180]
Angle 2 = [0,180]
Angle 3 = [0,360]

Noise in the XFEL diffraction patterns

The XFEL beam intensity is proxy for noise in the diffractions patterns (images)



Beam intensity

high

medium

low

low

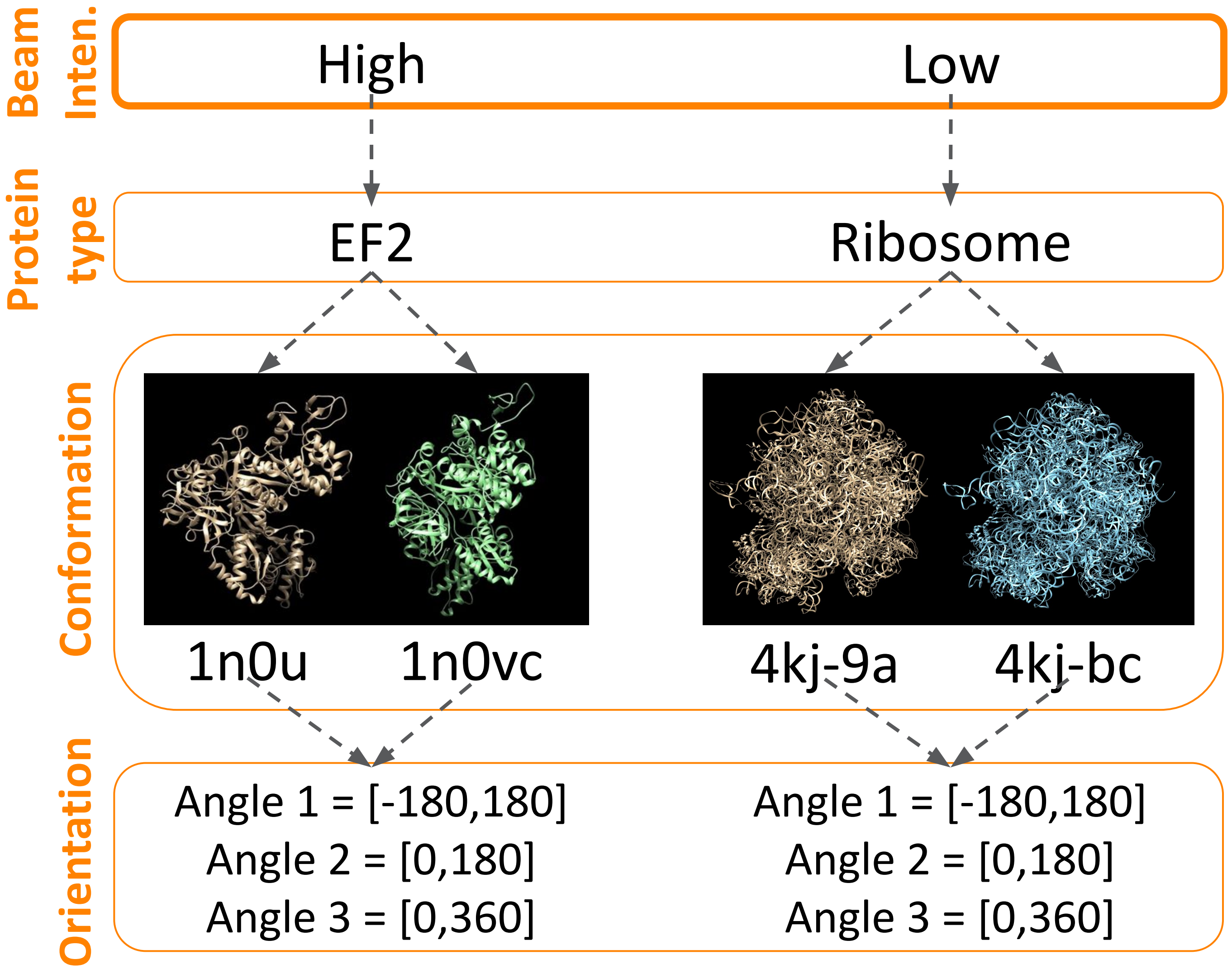
medium

high

Noise level

Lower beam intensity
→ Higher noise
→ Lower image resolution

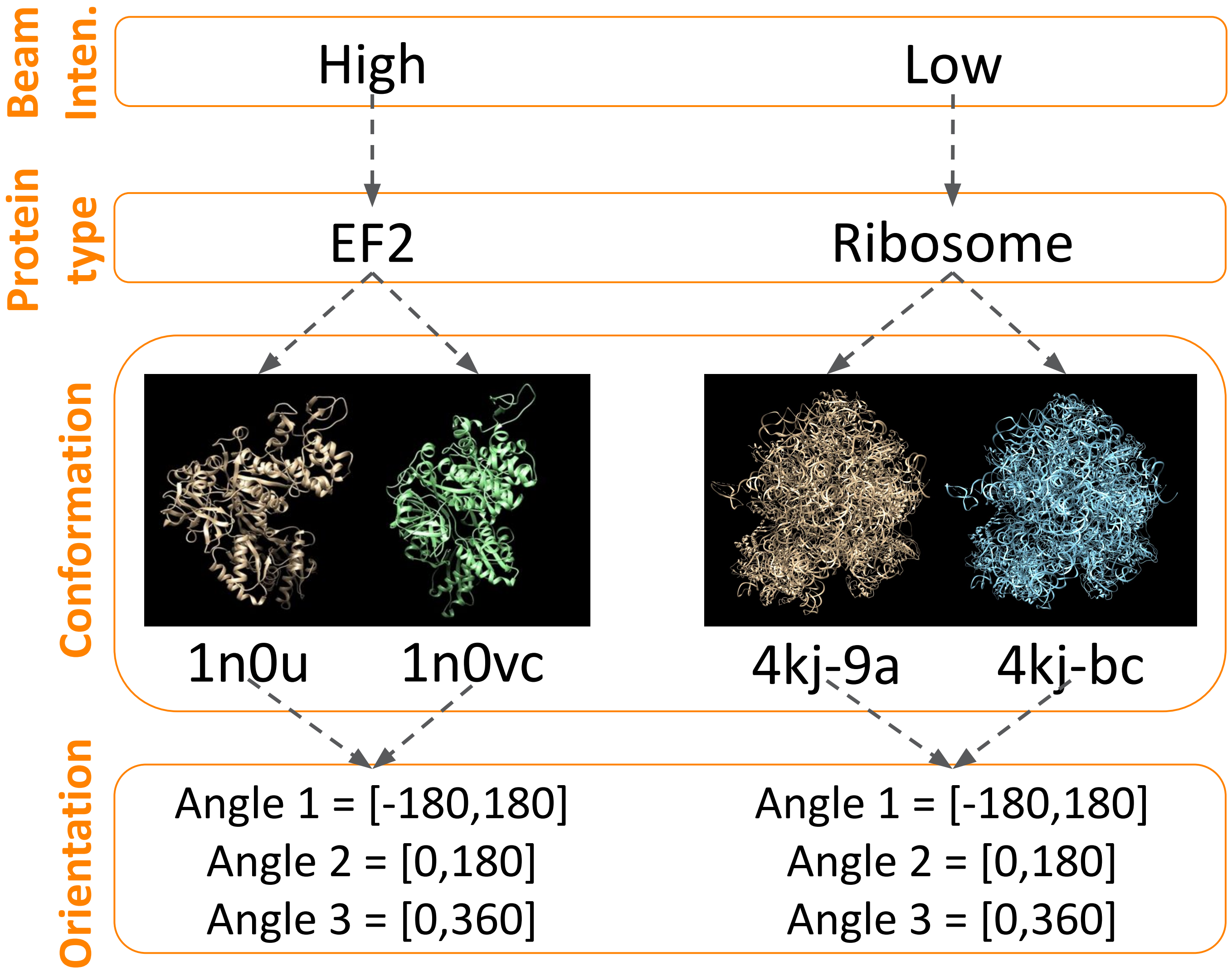
Identifying structural properties with XPSI



Identifying structural properties with XPSI

We quantify and validate XPSI’s ability to provide accurate structural properties predictions for **diverse datasets of diffraction patterns** (multiple orientations, conformations and protein types) with **different beam intensities**

- ➔ 10% testing data (~4000 diffraction patterns from each conformation)



Beam Inten.

High

Low

Protein type

EF2

Ribosome

Conformation

1n0u

1n0vc

4kj-9a

4kj-bc

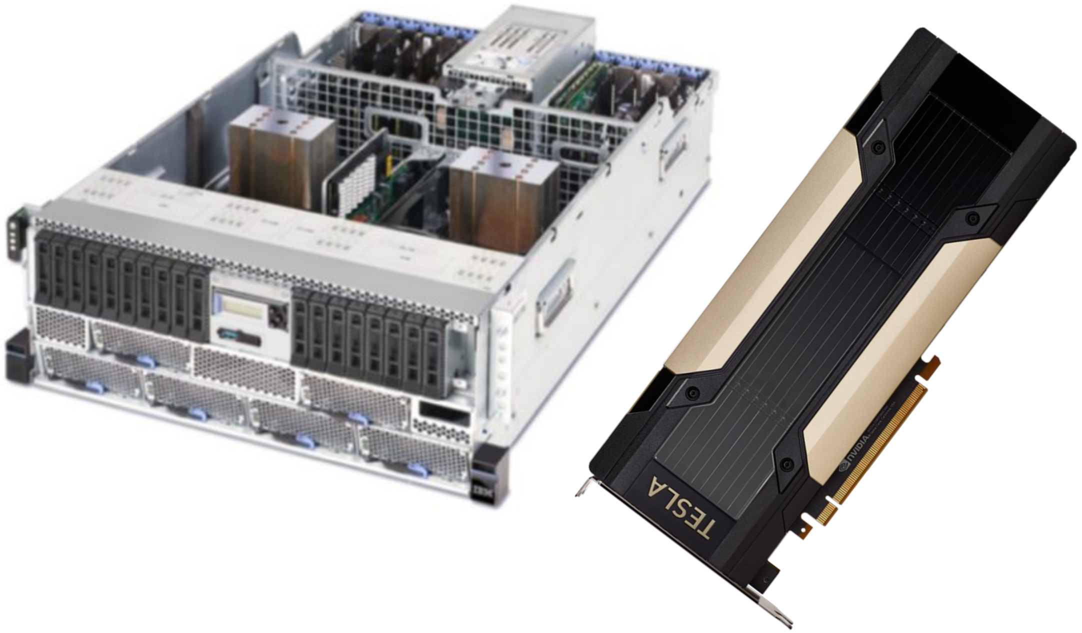
Orientation

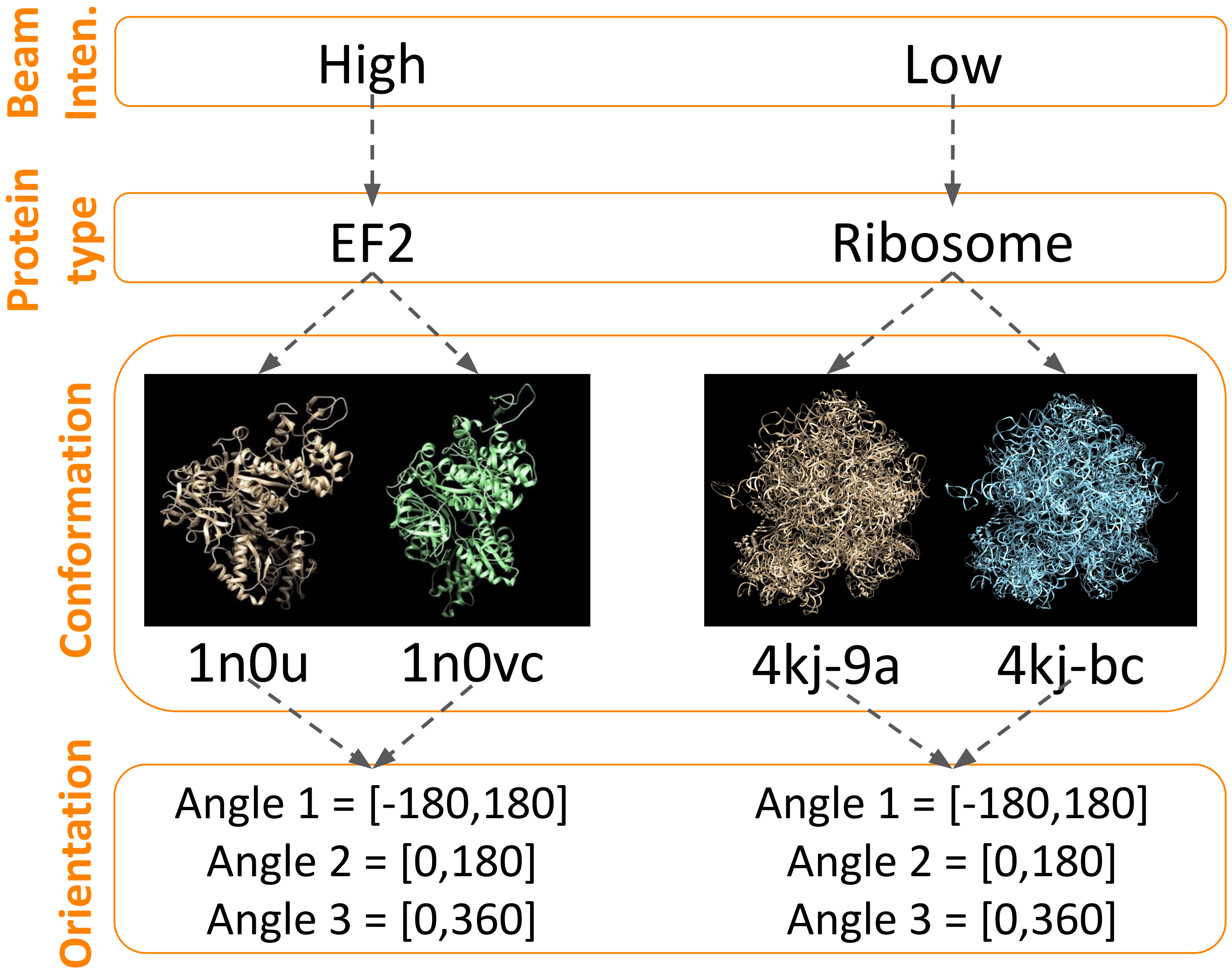
Angle 1 = [-180,180]
Angle 2 = [0,180]
Angle 3 = [0,360]

Angle 1 = [-180,180]
Angle 2 = [0,180]
Angle 3 = [0,360]

Computer infrastructure

1 x 32-core Power9 node (128 GB RAM) with 1 x GPU Nvidia V100s





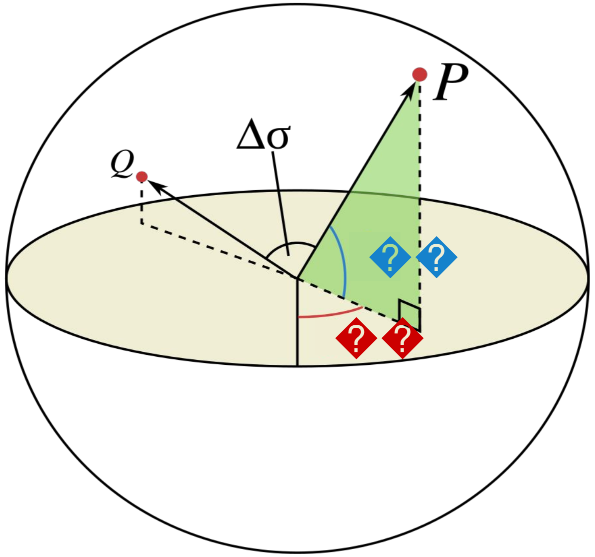
Orientation error

We measure the error to predict the three angles using two metrics

1. **Error degree**

The distance between two points on a sphere given Φ (Azimuth) and Θ (Altitude)

$$2\sqrt{\sin^2\left(\frac{\theta_2-\theta_1}{2}\right)+\cos(\theta_1)\cos(\theta_2)\sin^2\left(\frac{\phi_2-\phi_1}{2}\right)}$$

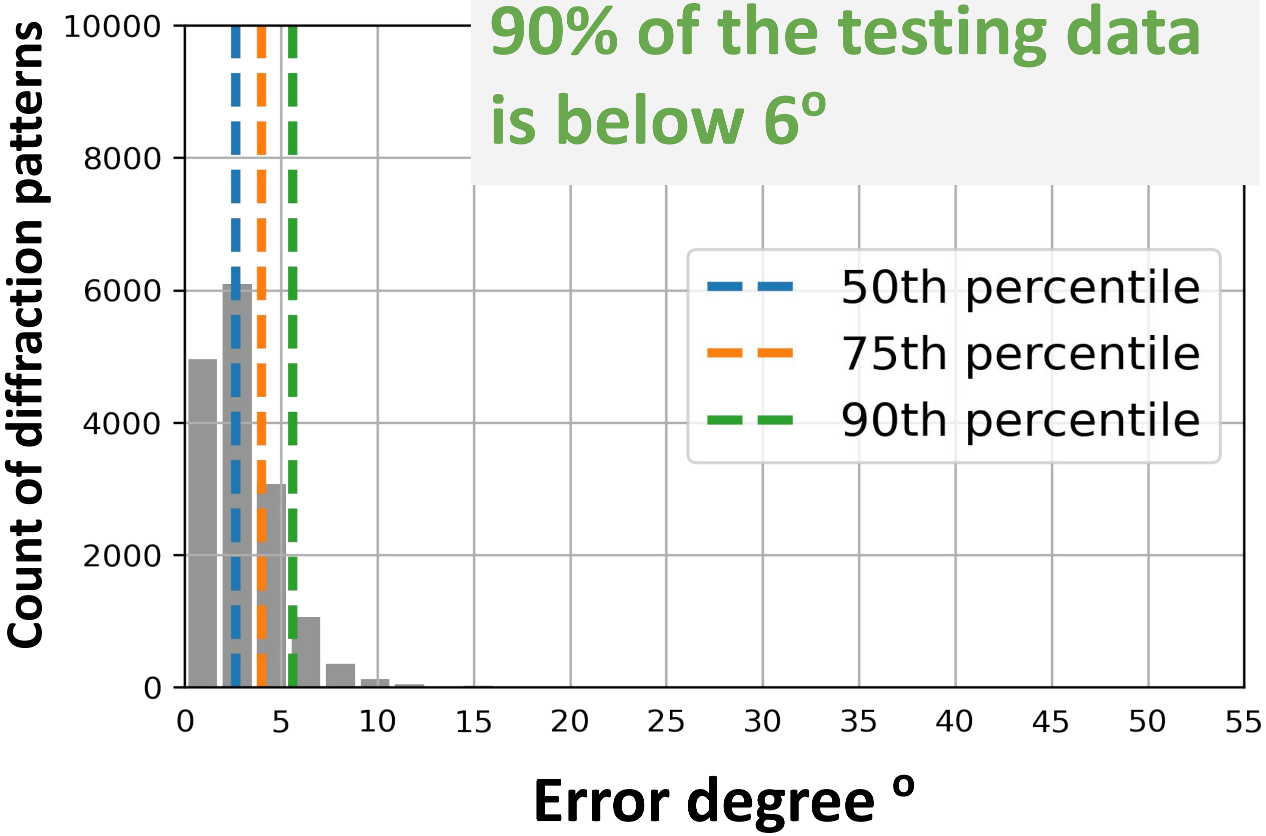


2. **Psi difference**

The difference between real and predicted Psi (Ψ) angle

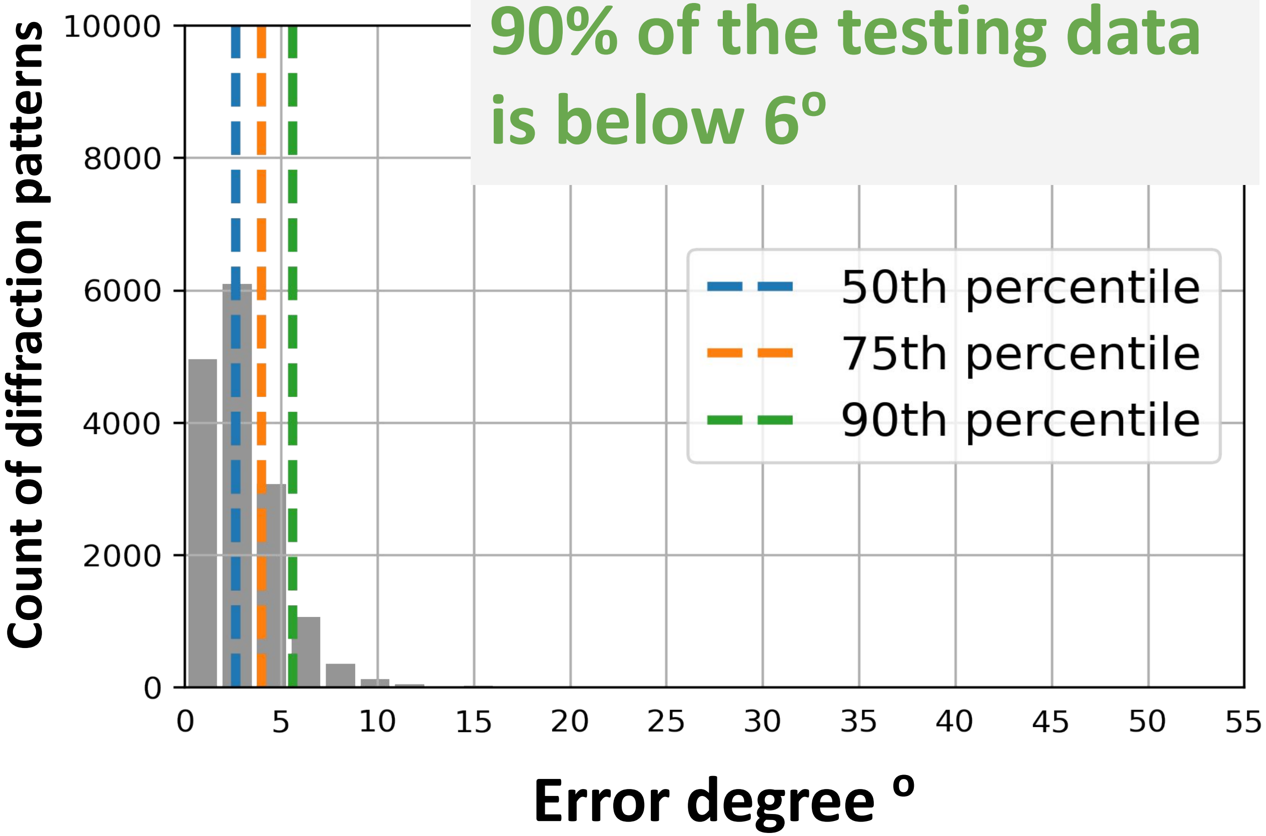
$$\psi_{real}-\psi_{predicted}$$

Orientation error

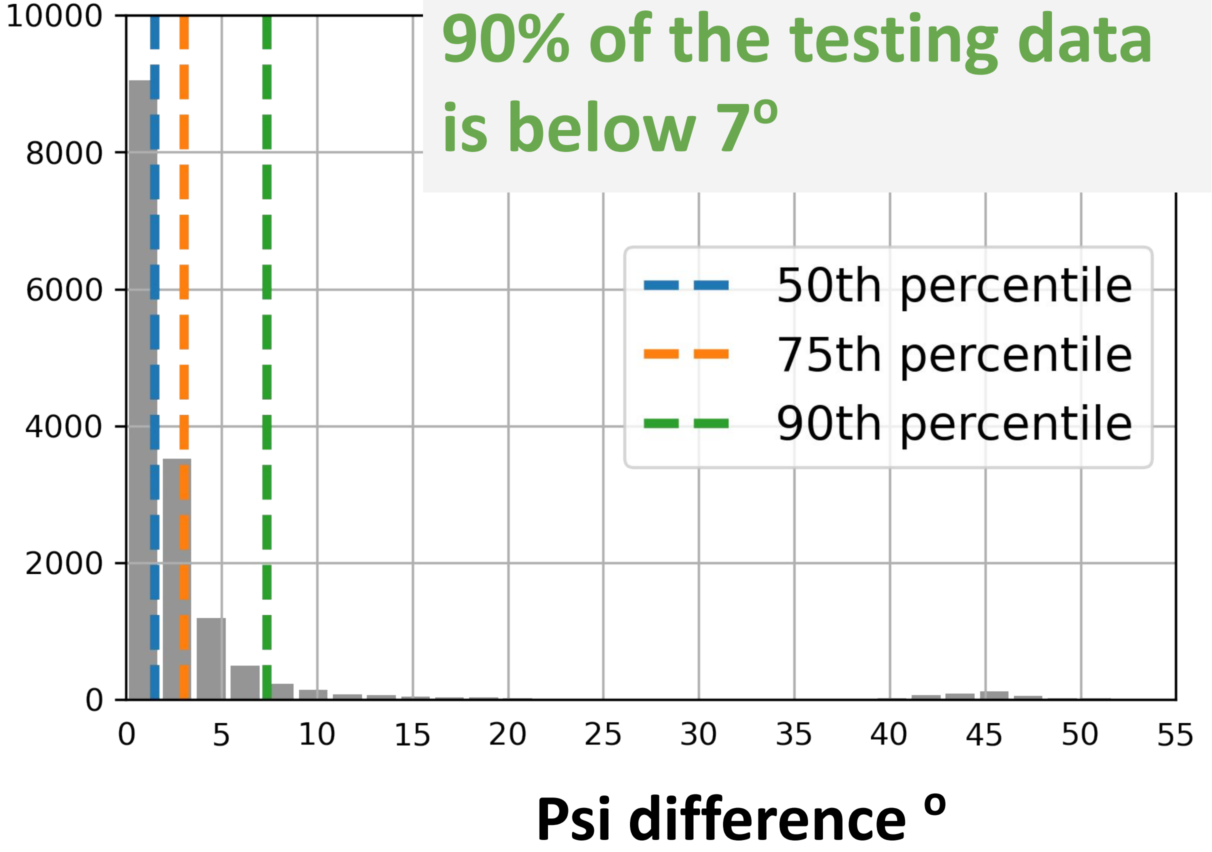


The error degree for 90% of the testing data is below 6°

Orientation error



The error degree for 90% of the testing data is below 6°



The psi difference for 90% of the testing data is below 7°

Conformation accuracy

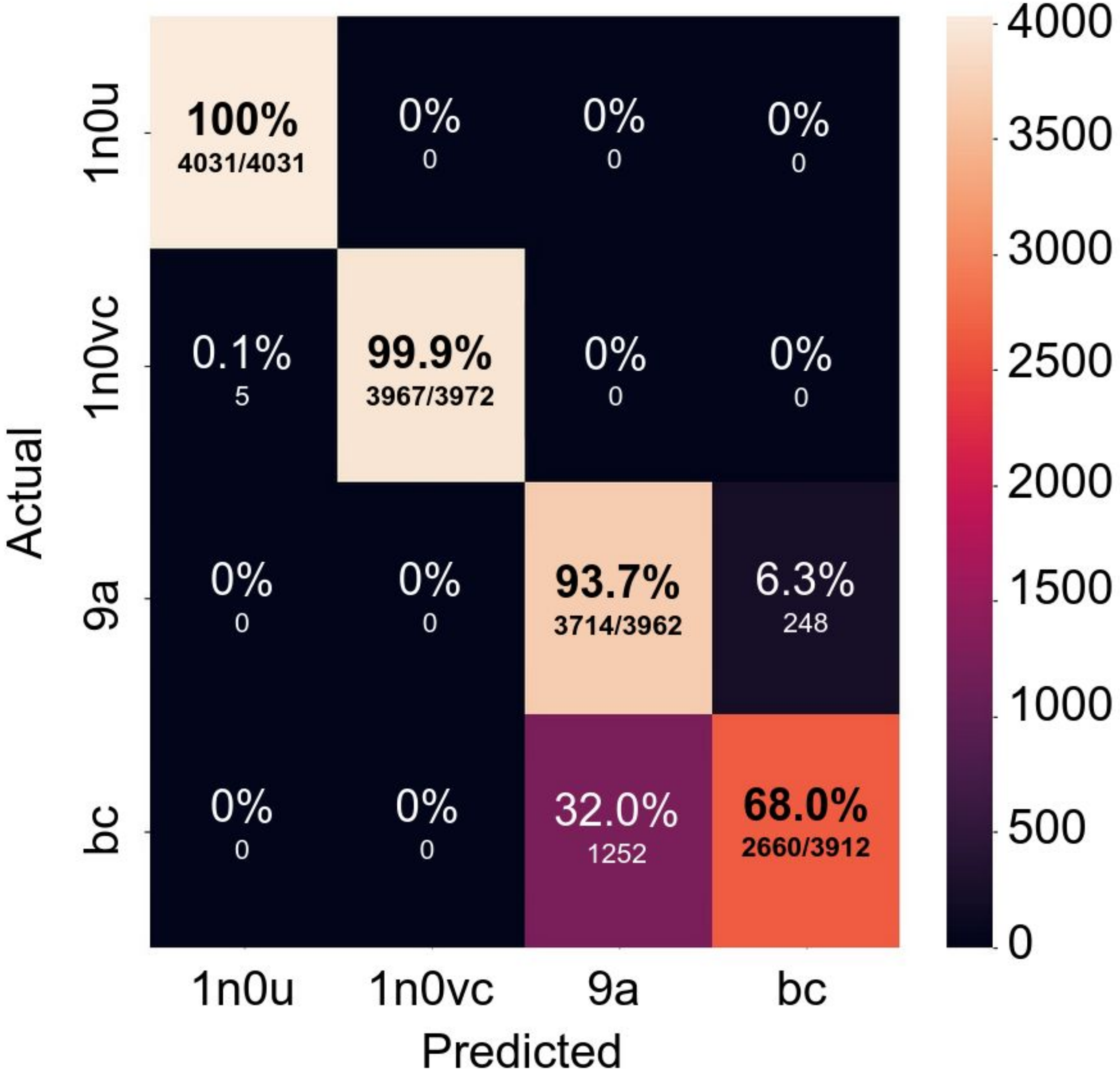
Accuracy: Represents the ratio of correct predictions over the total number of cases examined

$$\frac{TP + TN}{TP + TN + FP + FN}$$

Conformation accuracy

Accuracy: Represents the ratio of correct predictions over the total number of cases examined

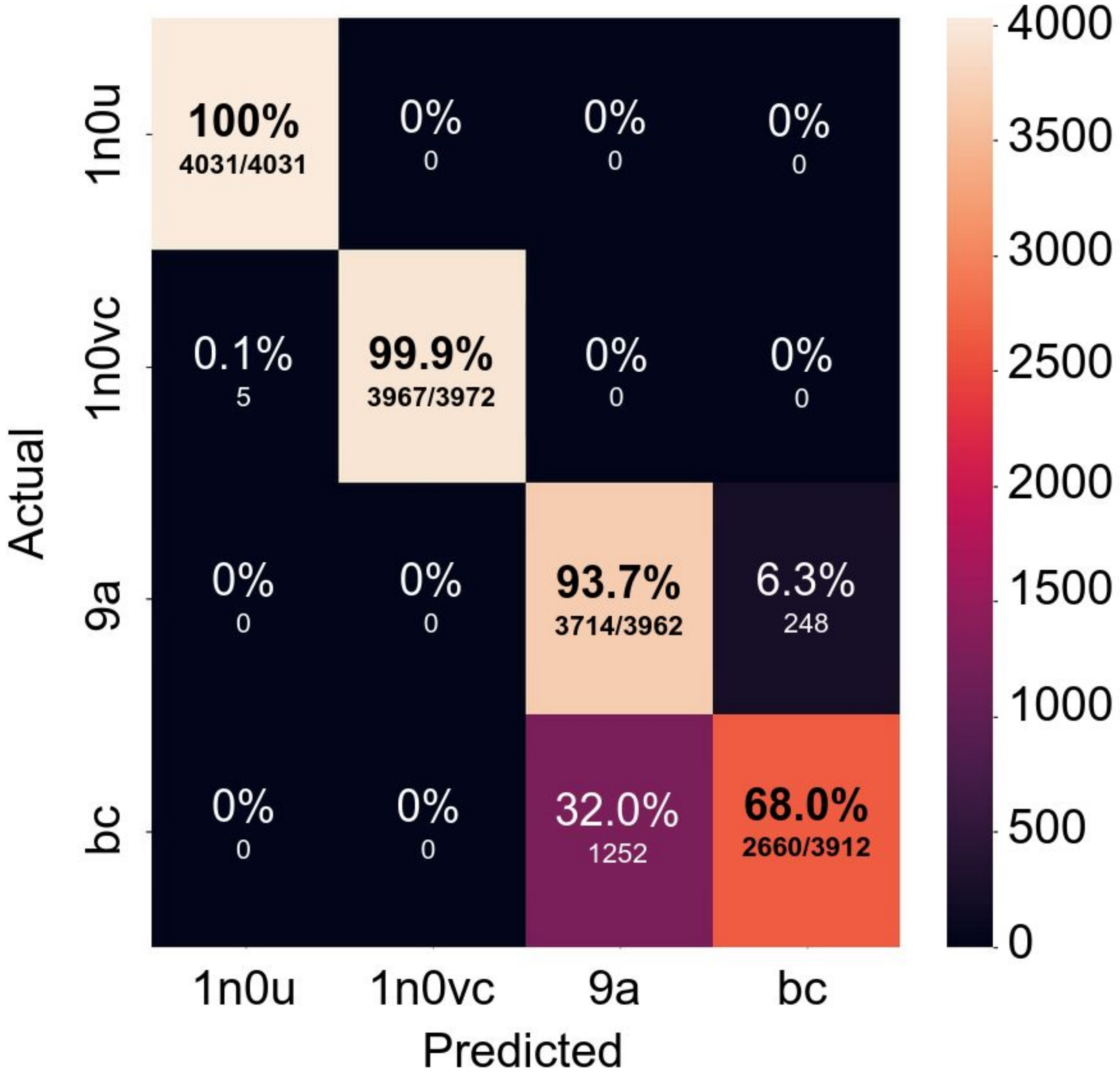
$$\frac{TP+TN}{TP+TN+FP+FN}$$



Actual \ Predicted	1n0u	1n0vc	9a	bc
1n0u	100% 4031/4031	0% 0	0% 0	0% 0
1n0vc	0.1% 5	99.9% 3967/3972	0% 0	0% 0
9a	0% 0	0% 0	93.7% 3714/3962	6.3% 248
bc	0% 0	0% 0	32.0% 1252	68.0% 2660/3912

Conformation accuracy

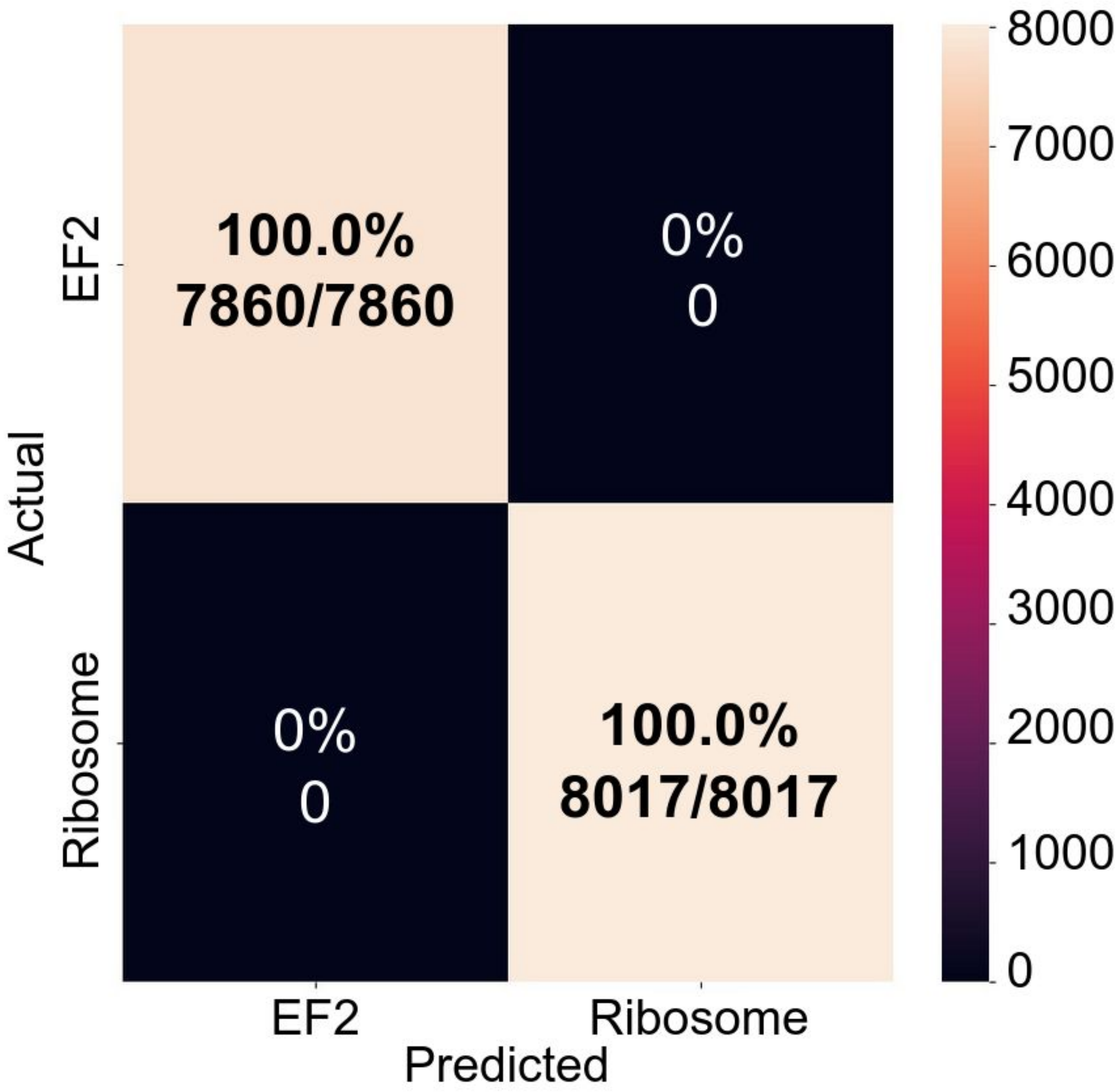
- **XPSI predicts** between 4 different **conformations with an accuracy of 90%** on average
- XPSI always predicts the conformations within the proteins (no inter-error class)



Actual \ Predicted	1n0u	1n0vc	9a	bc
1n0u	100% 4031/4031	0% 0	0% 0	0% 0
1n0vc	0.1% 5	99.9% 3967/3972	0% 0	0% 0
9a	0% 0	0% 0	93.7% 3714/3962	6.3% 248
bc	0% 0	0% 0	32.0% 1252	68.0% 2660/3912

Protein type accuracy

- **XPSI predicts** between 2 different **protein types with an accuracy of 100%**



XPSI remarks

We demonstrate the scientific robustness of XPSI in different challenges:

- **Identifying multiple proteins (100% of accuracy), conformations (90% of accuracy), and orientations (error degree < 6° and psi difference < 7°)**
- Differentiating between conformations (97% of accuracy) with similar, but not identical, structures of the same protein
- Identifying rotation in the diffraction patterns, even in the presence of symmetry (error degree < 10° and psi difference < 10°)

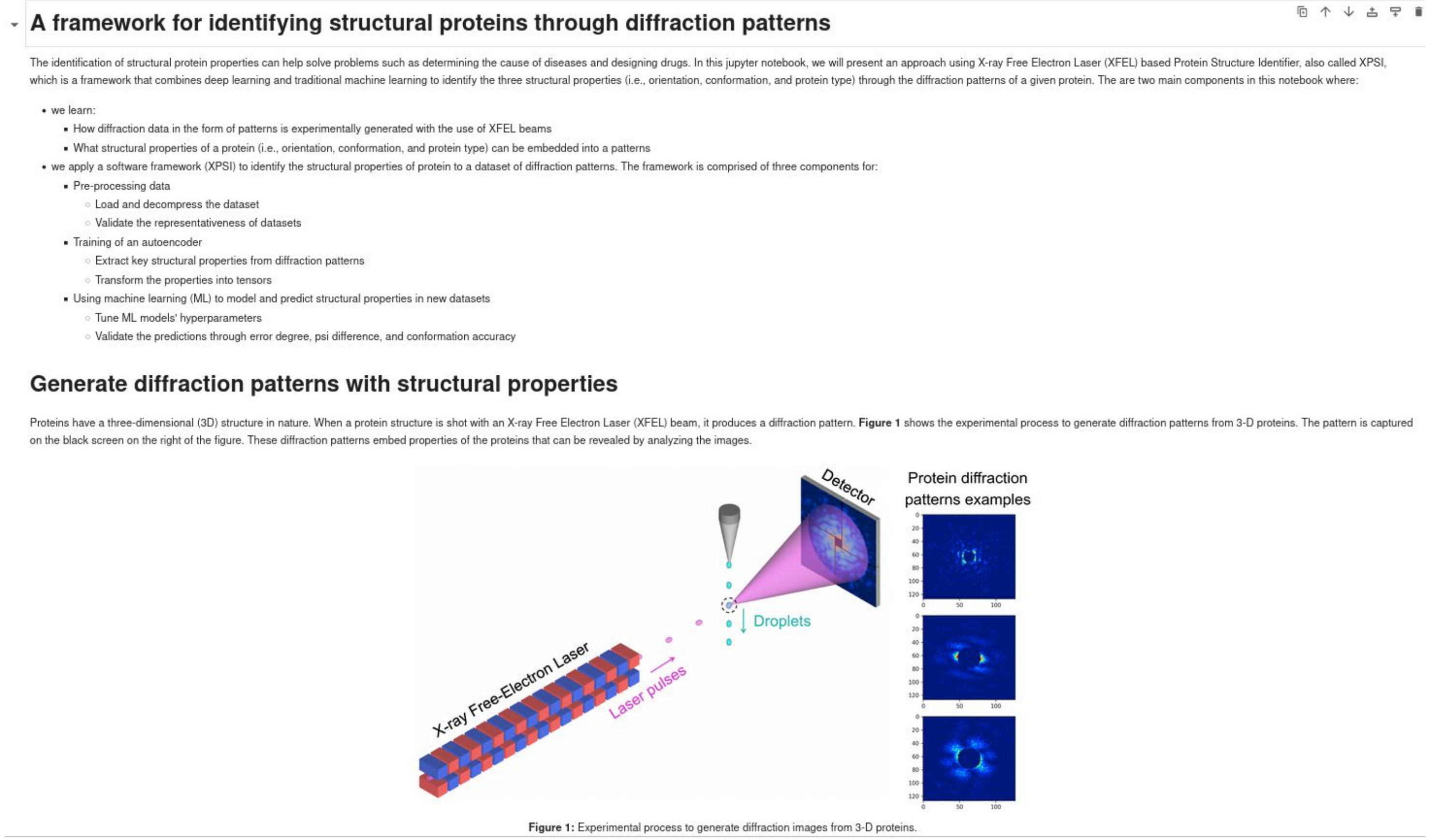
All of these capabilities are proven with different beam intensities. The lower the beam intensity the noisier the diffraction patterns, which affects the accuracy of the predictions

XPSI Jupyter notebook

We provide a Jupyter Notebook for shareability and portability of our framework

https://github.com/TauferLab/XPSI

SCAN ME



A framework for identifying structural proteins through diffraction patterns

The identification of structural protein properties can help solve problems such as determining the cause of diseases and designing drugs. In this jupyter notebook, we will present an approach using X-ray Free Electron Laser (XFEL) based Protein Structure Identifier, also called XPSI, which is a framework that combines deep learning and traditional machine learning to identify the three structural properties (i.e., orientation, conformation, and protein type) through the diffraction patterns of a given protein. The are two main components in this notebook where:

- we learn:
  - How diffraction data in the form of patterns is experimentally generated with the use of XFEL beams
  - What structural properties of a protein (i.e., orientation, conformation, and protein type) can be embedded into a patterns
- we apply a software framework (XPSI) to identify the structural properties of protein to a dataset of diffraction patterns. The framework is comprised of three components for:
  - Pre-processing data
    - Load and decompress the dataset
    - Validate the representativeness of datasets
  - Training of an autoencoder
    - Extract key structural properties from diffraction patterns
    - Transform the properties into tensors
  - Using machine learning (ML) to model and predict structural properties in new datasets
    - Tune ML models' hyperparameters
    - Validate the predictions through error degree, psi difference, and conformation accuracy

Generate diffraction patterns with structural properties

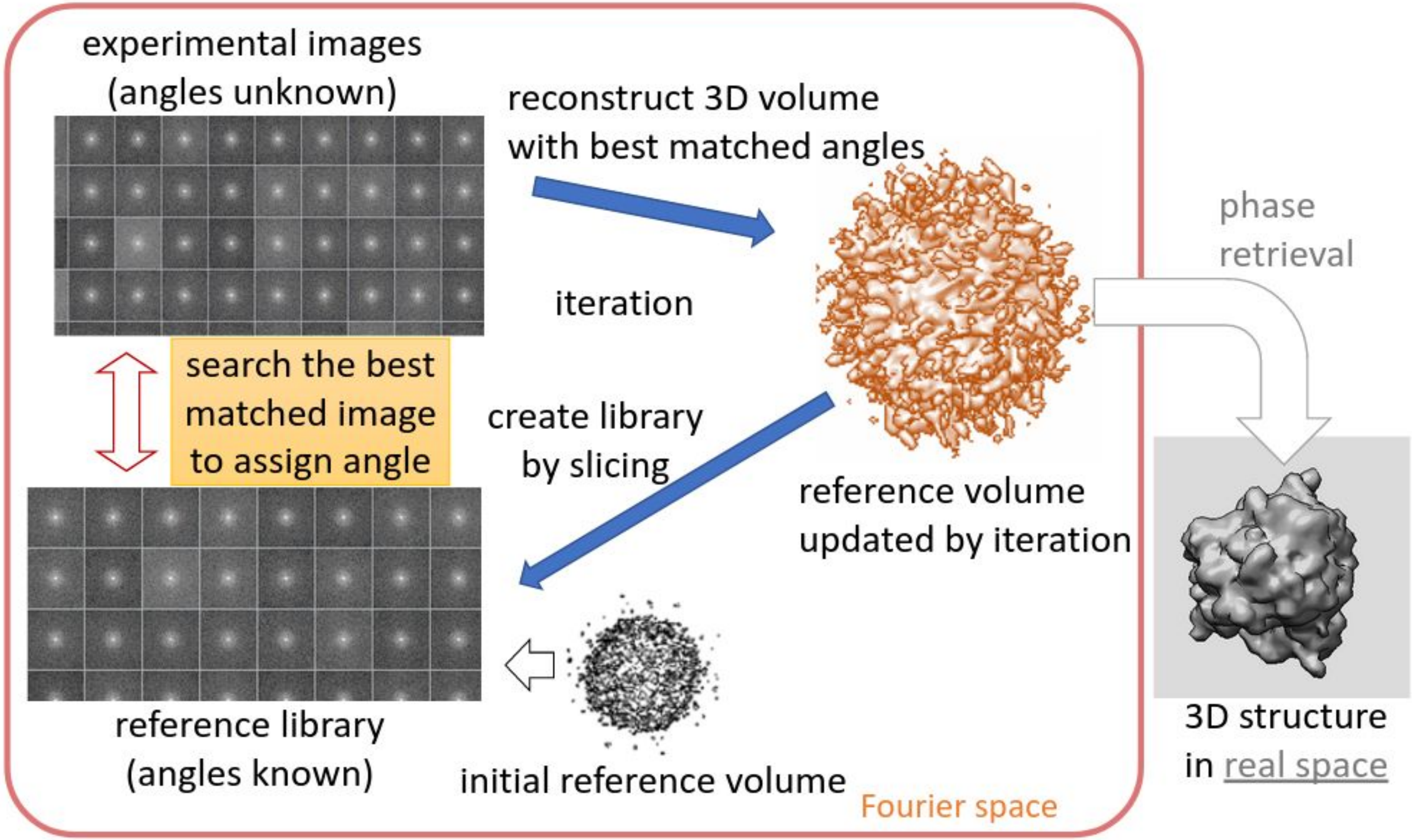
Proteins have a three-dimensional (3D) structure in nature. When a protein structure is shot with an X-ray Free Electron Laser (XFEL) beam, it produces a diffraction pattern. **Figure 1** shows the experimental process to generate diffraction patterns from 3-D proteins. The pattern is captured on the black screen on the right of the figure. These diffraction patterns embed properties of the proteins that can be revealed by analyzing the images.

Detector
Protein diffraction patterns examples
Droplets
X-ray Free-Electron Laser
Laser pulses

Figure 1: Experimental process to generate diffraction images from 3-D proteins.

XFEL slice matching for 3D reconstruction

Apply our XPSI framework to XFEL slice matching for 3D reconstruction



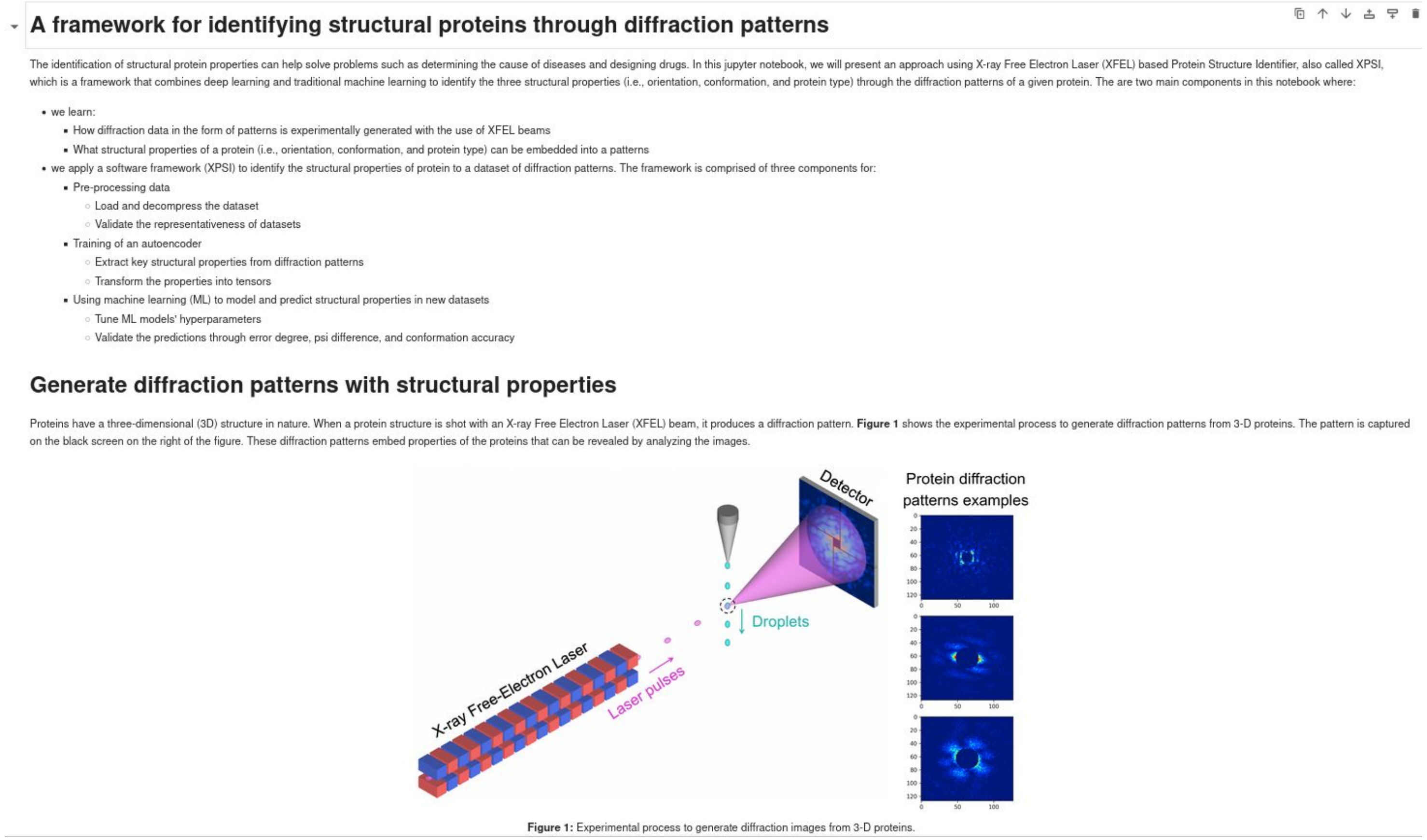
Nakano et al., J. Synchrotron Rad. (2017). 24, 727–737

XPSI Jupyter notebook

We provide a Jupyter Notebook for shareability and portability of our framework

https://github.com/TauferLab/XPSI

SCAN ME



A framework for identifying structural proteins through diffraction patterns

The identification of structural protein properties can help solve problems such as determining the cause of diseases and designing drugs. In this jupyter notebook, we will present an approach using X-ray Free Electron Laser (XFEL) based Protein Structure Identifier, also called XPSI, which is a framework that combines deep learning and traditional machine learning to identify the three structural properties (i.e., orientation, conformation, and protein type) through the diffraction patterns of a given protein. The are two main components in this notebook where:

- we learn:
  - How diffraction data in the form of patterns is experimentally generated with the use of XFEL beams
  - What structural properties of a protein (i.e., orientation, conformation, and protein type) can be embedded into a patterns
- we apply a software framework (XPSI) to identify the structural properties of protein to a dataset of diffraction patterns. The framework is comprised of three components for:
  - Pre-processing data
    - Load and decompress the dataset
    - Validate the representativeness of datasets
  - Training of an autoencoder
    - Extract key structural properties from diffraction patterns
    - Transform the properties into tensors
  - Using machine learning (ML) to model and predict structural properties in new datasets
    - Tune ML models' hyperparameters
    - Validate the predictions through error degree, psi difference, and conformation accuracy

Generate diffraction patterns with structural properties

Proteins have a three-dimensional (3D) structure in nature. When a protein structure is shot with an X-ray Free Electron Laser (XFEL) beam, it produces a diffraction pattern. **Figure 1** shows the experimental process to generate diffraction patterns from 3-D proteins. The pattern is captured on the black screen on the right of the figure. These diffraction patterns embed properties of the proteins that can be revealed by analyzing the images.

Detector
Protein diffraction patterns examples
Droplets
X-ray Free-Electron Laser
Laser pulses

Figure 1: Experimental process to generate diffraction images from 3-D proteins.