



LUND
UNIVERSITY

Protein Superimposition with Machine Learning and Deep Learning

Students: Theodora Gaiceanu, Diego Figueroa
Supervisor: Daniel Varela





INTRODUCTION

Why this project?

- Use DL and ML to align 2 proteins
- Input: original protein + rotated one
- Output:
 - Rotation values or,
 - Zeal score

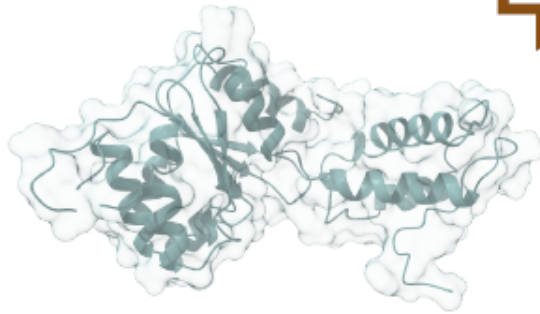
- The shape of a protein - an essential role in drug discovery
- Aligning proteins - function & relationships between proteins
- Increasing dataset - ANN model

Goal of the project

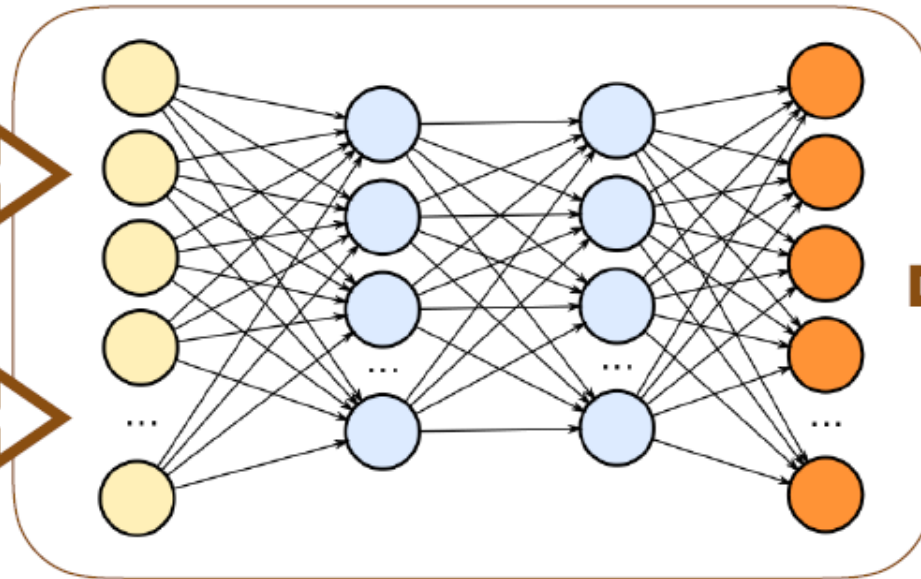
Input 1



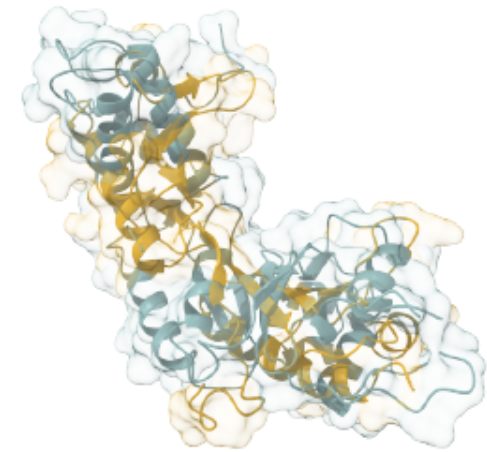
Input 2



Deep Learning method



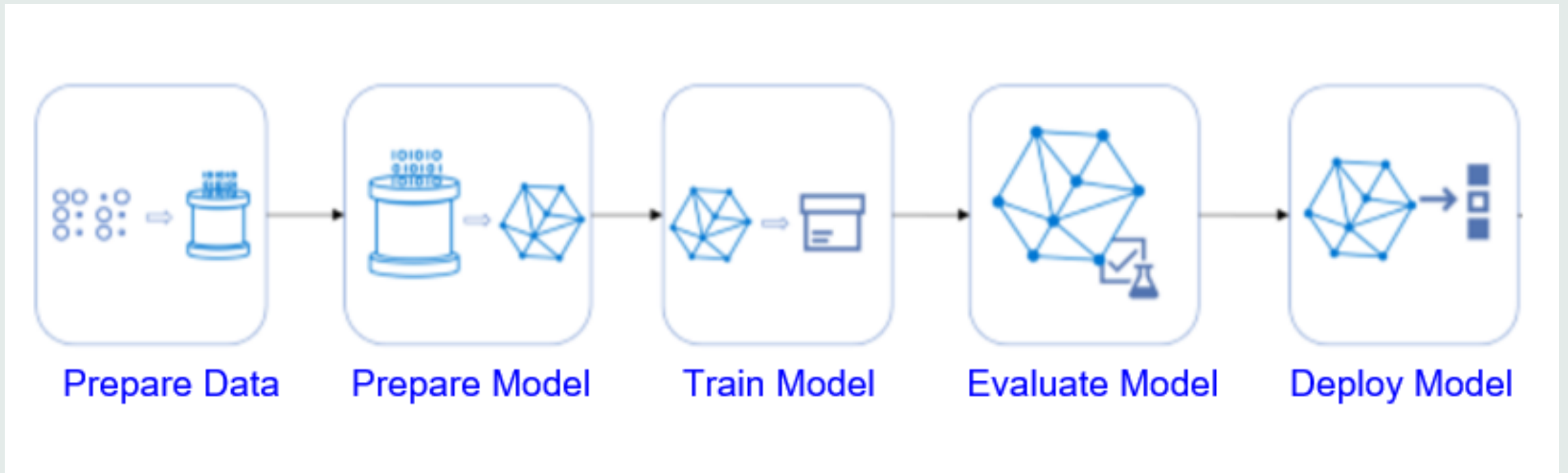
Aligned Proteins





WORKFLOW

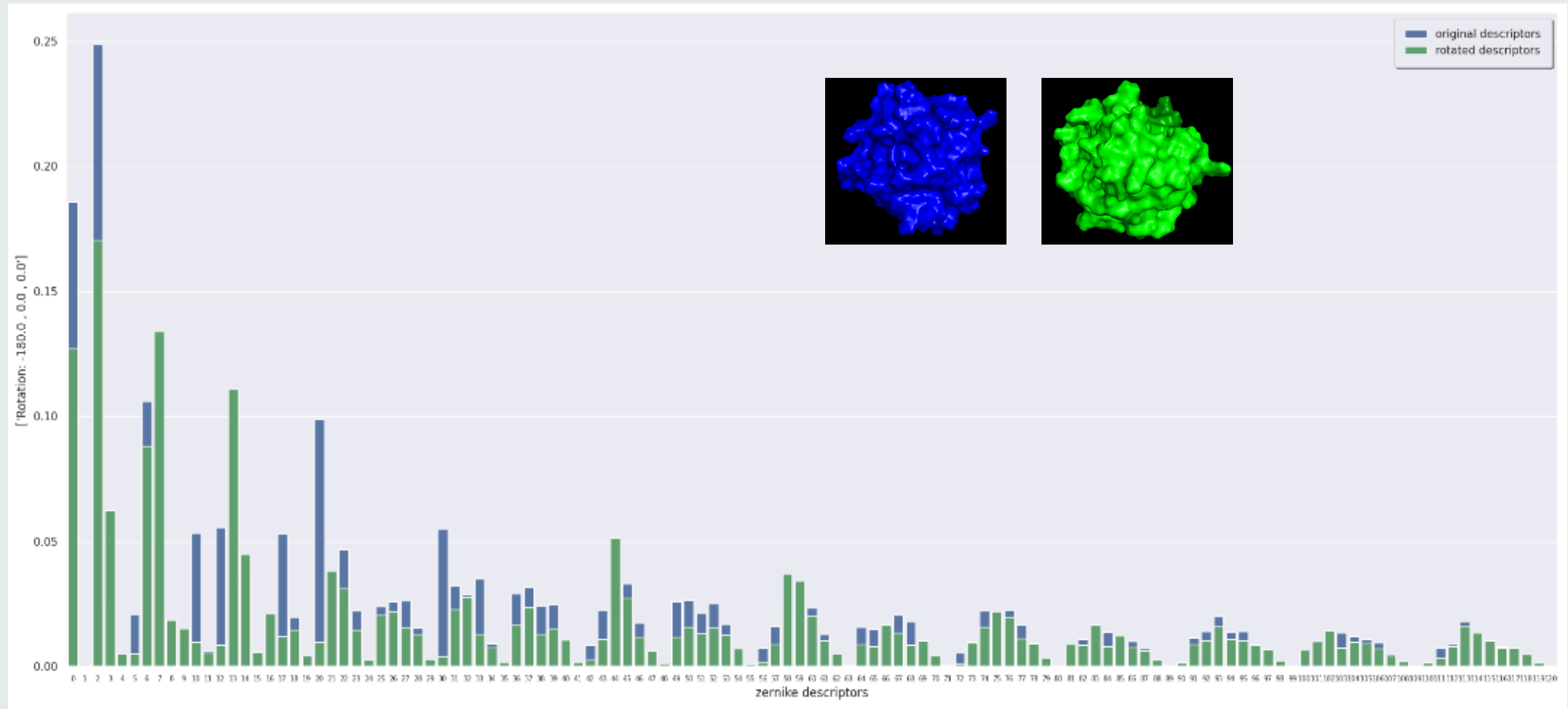
Workflow of the project

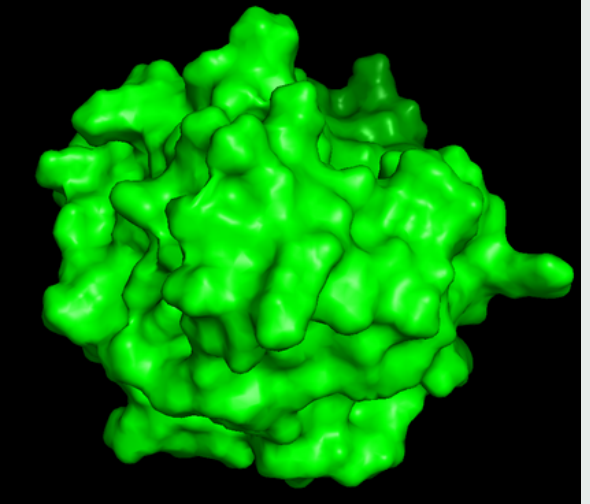
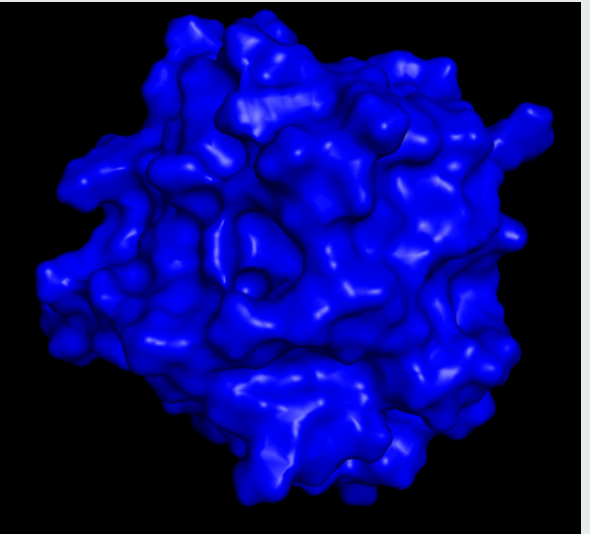
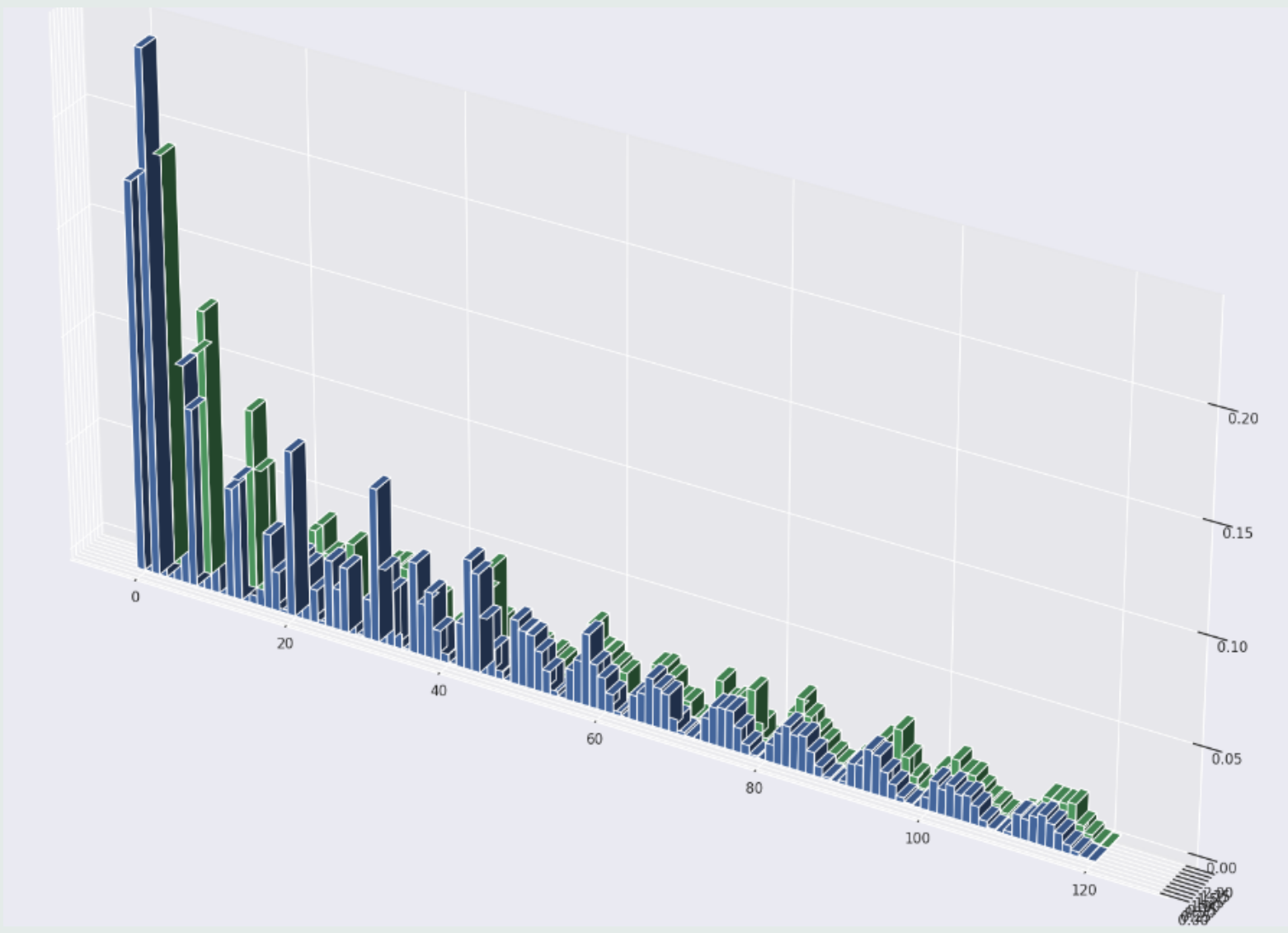




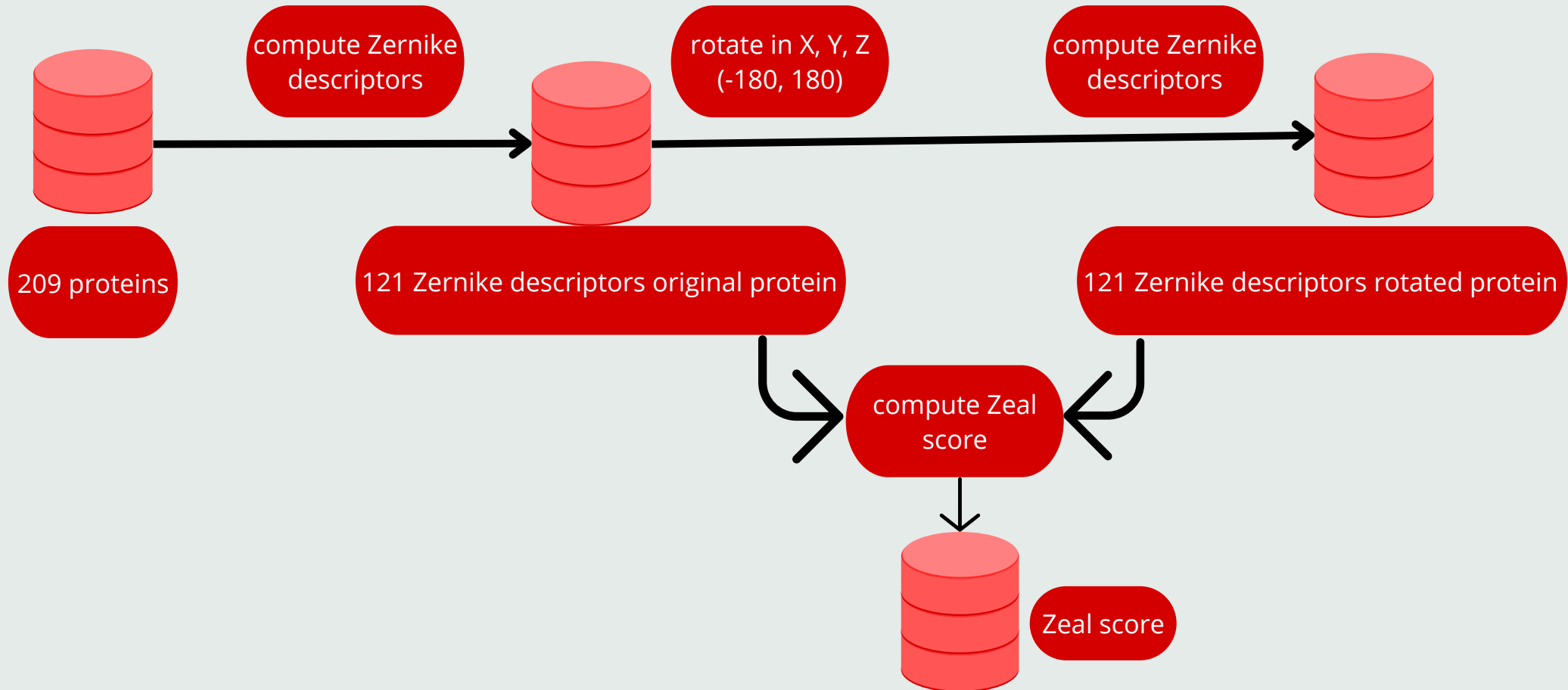
DATA PREPARATION & PREPROCESSING

Zernike descriptors

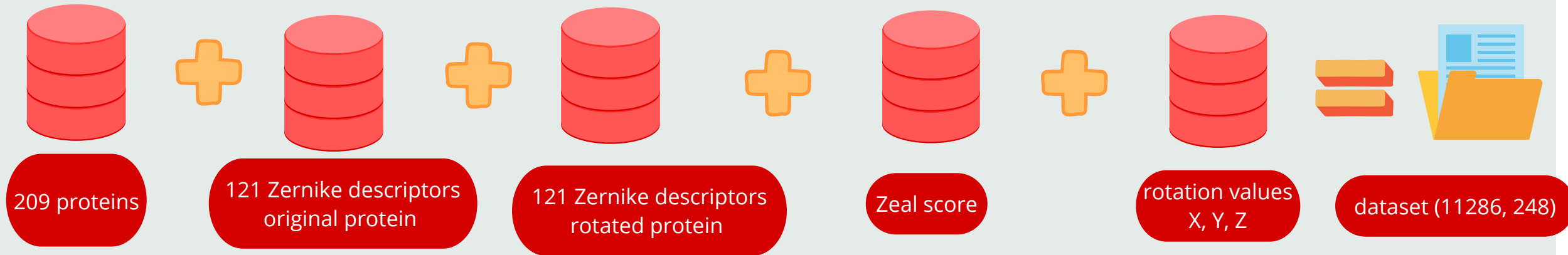




Data preparation



Data preparation



Preprocessing:

- scale Zernike descriptors - MinMax scaler
- check null values

Dataset

0	pdb_ID	zd_I_0	zd_I_1	zd_I_2	...	zd_R_120	r_0	r_1	r_2	zeal
0	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000088	-180.0	0.0	0.0	0.603379
1	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000042	0.0	-180.0	0.0	0.719376
2	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000176	0.0	0.0	-180.0	0.570384
3	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000083	-160.0	0.0	0.0	0.643067
4	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000039	0.0	-160.0	0.0	0.702196
5	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000157	0.0	0.0	-160.0	0.583470
6	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000173	-140.0	0.0	0.0	0.672413
7	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000034	0.0	-140.0	0.0	0.691126
8	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000158	0.0	0.0	-140.0	0.590590
9	d3fv3g_.ent	0.180773	3.328894e-17	0.242312	...	0.000226	-120.0	0.0	0.0	0.707422



MODELS

Models

- 80% training + 20% testing
- First approach
 - Input: zernike descriptors original protein + zernike descriptors rotated protein
 - Output: rotation values
- Second approach
 - Input: zernike descriptors original protein + zernike descriptors rotated protein + rotation values
 - Output: zeal score

Models

- Considered models:
 - Multioutput Random Forest Regressor
 - Multioutput MLP Regressor
 - CNN Regressor (three outputs)

Multioutput Random Forest

Multioutput Random Forest Regressor - output - rotation values:

- 100 trees
- MSE: 4486.485

Models - Multioutput MLP Regressor

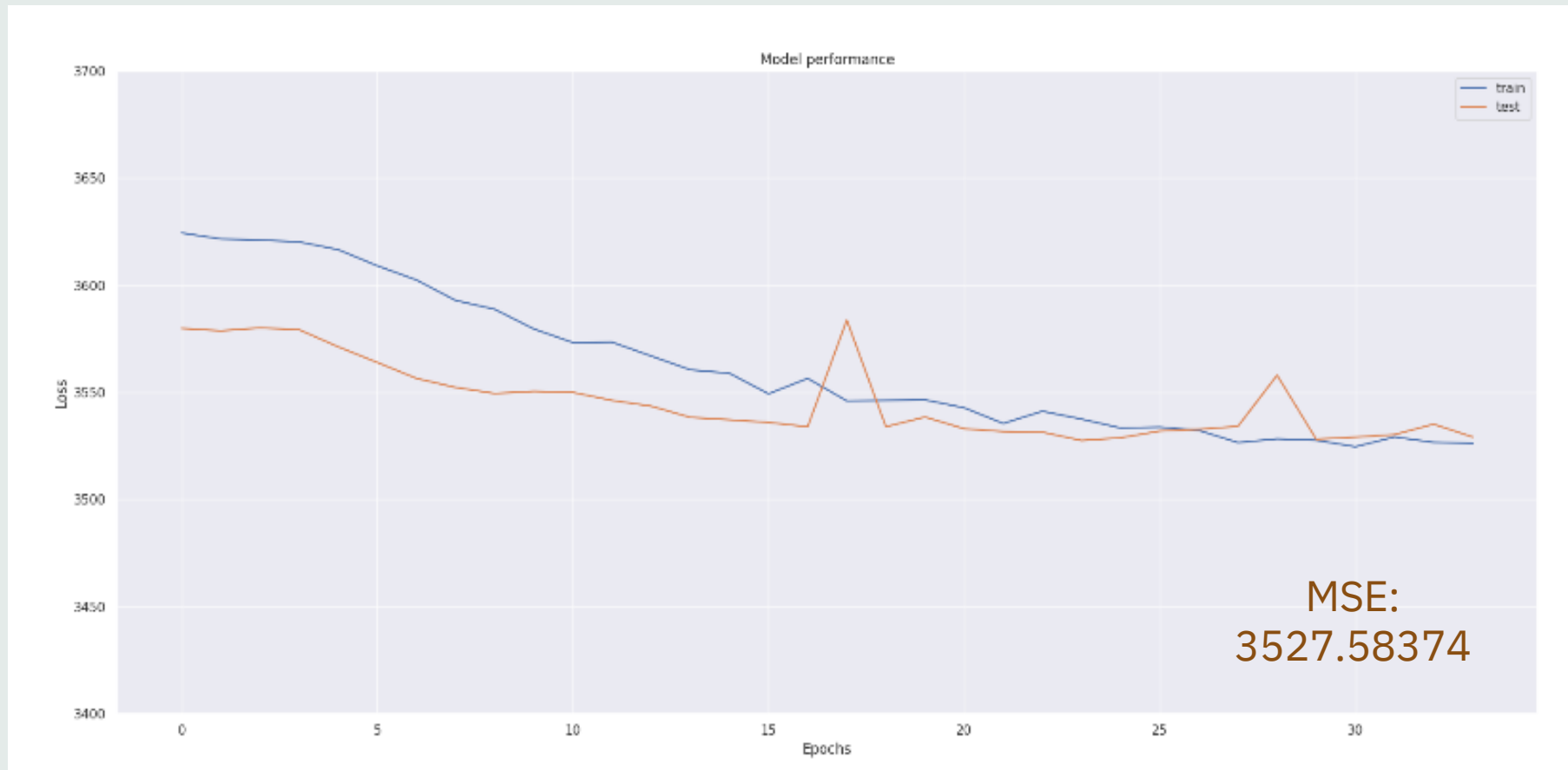
- Multioutput MLP Regressor - output - rotation values
 - 43 hidden layers
 - 200 max iterations
 - Early stopping

- MSE: 3578.577

Models - CNN Regressor

- CNN Regressor - output - rotation values
 - 2 1D convolutional layers
 - Flatten layer
 - Dense layer
 - Output layer
 - 200 max epochs
 - Early stopping

CNN Regressor



Loss function (MSE) for the CNN Regressor (output rotation values)

Model: "sequential"

Layer (type)	Output Shape	Param #
reshape (Reshape)	(None, 242, 1)	0
conv1d (Conv1D)	(None, 238, 64)	384
max_pooling1d (MaxPooling1D)	(None, 79, 64)	0
dropout (Dropout)	(None, 79, 64)	0
conv1d_1 (Conv1D)	(None, 77, 32)	6176
max_pooling1d_1 (MaxPooling1D)	(None, 25, 32)	0
dropout_1 (Dropout)	(None, 25, 32)	0
flatten (Flatten)	(None, 800)	0
dense (Dense)	(None, 64)	51264
dropout_2 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 3)	195

=====
Total params: 58,019
Trainable params: 58,019
Non-trainable params: 0

Models - CNN Regressor

CNN Regressor (output: rotation values) model
configuration

Models

- Considered models:
 - Random Forest Regressor
 - MLP Regressor
 - CNN Regressor (one output)

Random Forest

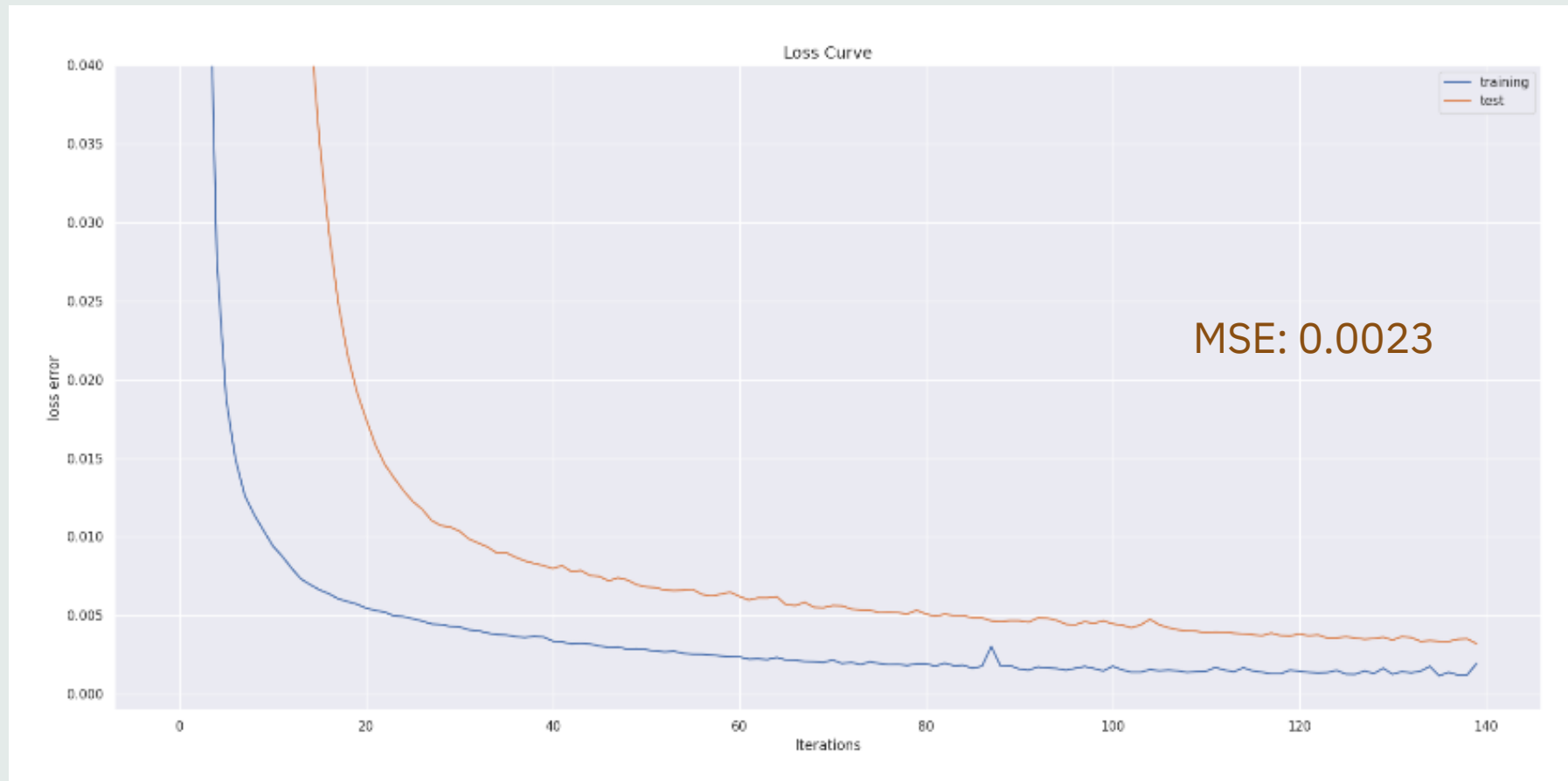
Random Forest Regressor - output - Zeal score:

- 100 trees
- MSE: 0.0016

Models - MLP Regressor

- MLP Regressor - output -Zeal score
 - 43 hidden layers
 - 200 max iterations
 - Early stopping

MLP Regressor



Loss function (MSE) for the MLP Regressor (output Zeal score)

Models - CNN Regressor

- CNN Regressor - output - Zeal score
 - 3 1D convolutional layers
 - Flatten layer
 - Dense layer
 - Output layer
 - 100 max epochs
 - Early stopping

Model: "sequential"

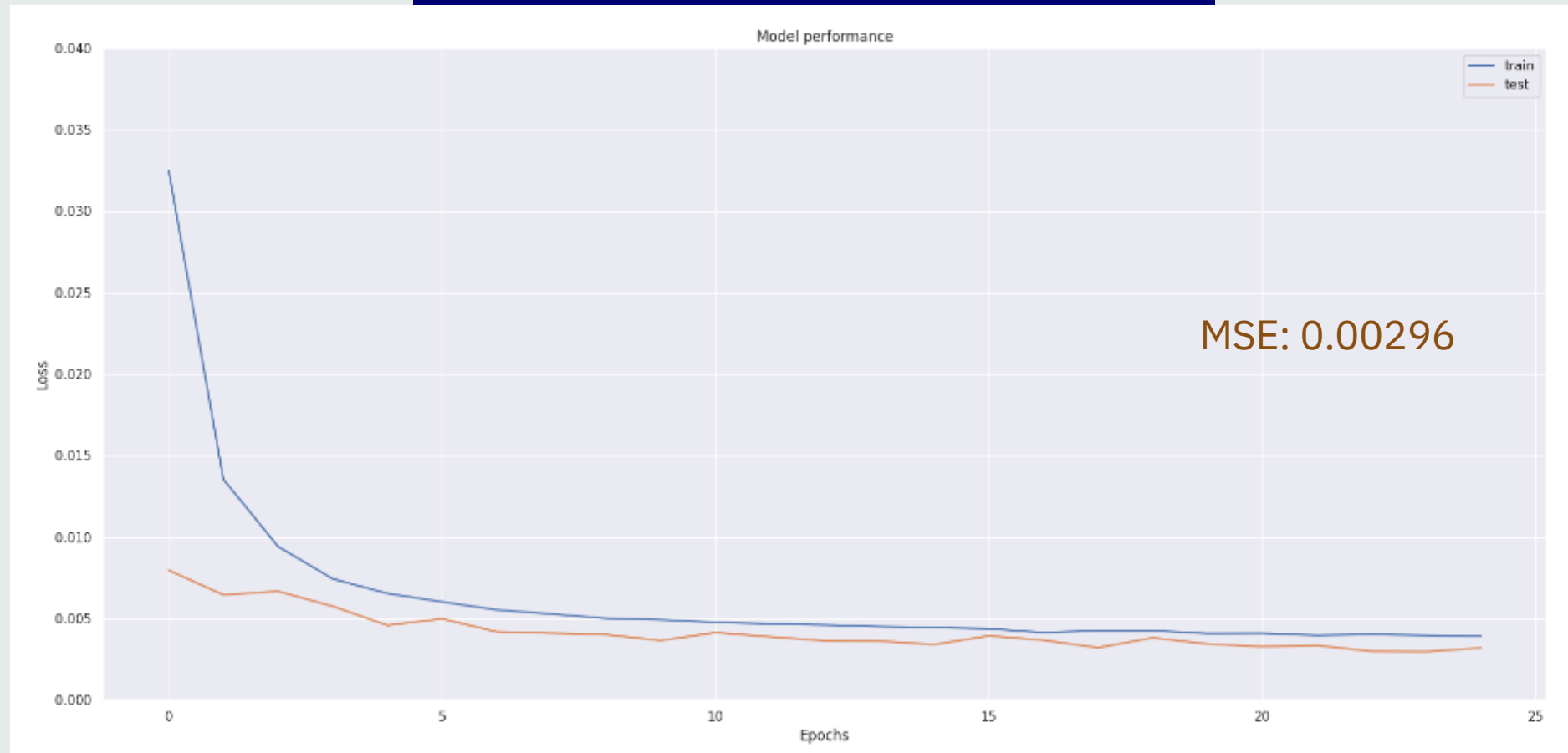
Layer (type)	Output Shape	Param #
reshape (Reshape)	(None, 245, 1)	0
conv1d (Conv1D)	(None, 241, 64)	384
max_pooling1d (MaxPooling1D)	(None, 80, 64)	0
dropout (Dropout)	(None, 80, 64)	0
conv1d_1 (Conv1D)	(None, 78, 32)	6176
max_pooling1d_1 (MaxPooling1D)	(None, 26, 32)	0
dropout_1 (Dropout)	(None, 26, 32)	0
conv1d_2 (Conv1D)	(None, 24, 16)	1552
flatten (Flatten)	(None, 384)	0
dense (Dense)	(None, 64)	24640
dropout_2 (Dropout)	(None, 64)	0
dense_1 (Dense)	(None, 1)	65

=====
Total params: 32,817
Trainable params: 32,817
Non-trainable params: 0

Models - CNN Regressor

CNN Regressor (output: Zeal score) model configuration

CNN Regressor

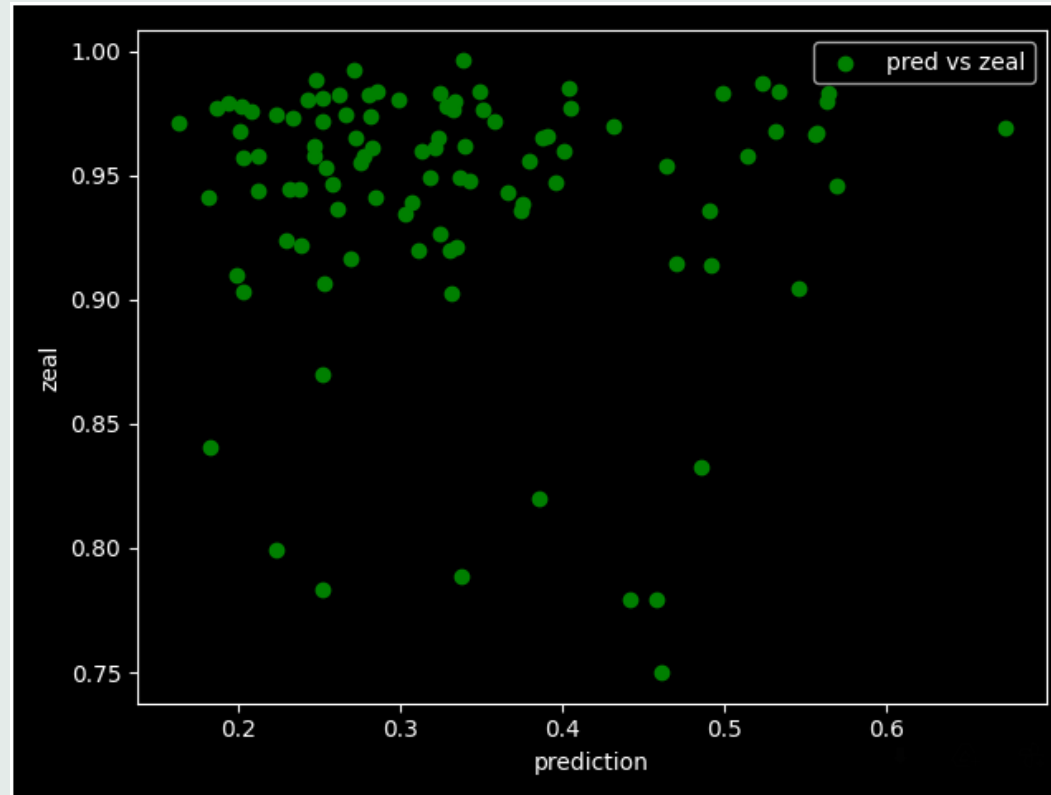


Loss function (MSE) for the CNN Regressor (output Zeal score)



DEPLOYMENT

Validating in real case



Performance of the models with real data



CONCLUSIONS

Conclusions

- data generation - computational expensive, tedious process
- more data is needed
- the models outputting the Zeal score work better
- 2D convolutional layers - possible improvement

THANK
YOU



LUND
UNIVERSITY