Vilnius University

Mathematics and Informatics Faculty

Institute of Informatics

Bioinformatics study program

Protein thermostability prediction using sequence representations from protein language models

Author: Ieva Pudžiuvelytė

Supervisor: Kliment Olechnovič, PhD

Course work project

Vilnius, 2022

Contents

1 Introduction

This work is a prolongation of the previous work - the model that performed binary classification into thermostability classes. The model took ESM-1b protein embeddings as input and provided prediction for each protein, how likely it belongs to the thermostable class.

2 Abstract in Lithuanian (Santrauka)

3 Theory

3.1 ESM-1b embeddings

3.2 ProtTrans embeddings

4 Methods

4.1 Objective of this work

The main objective of this work is to analyse which numerical representation of proteins is the most suitable to use as input for the neural network model to make binary protein classification into thermostability classes. Additionally, it was decided to try model architectures with one or two hidden layers and evaluate whether the different architecture improves the performance.

4.2 Data set

The data set, which was used for the research of principal components' usage as input this work, was used the same as for the training, validation, and testing of the single-layer perceptron (SLP) with mean ESM-1b representations in the previous work. Although, for the research of other representations, the data set was filtered out of identical sequences to get more accurate evaluations.

Table 1: Number of sequences with embeddings before and after filtering the data set

Subset	Original	After filtering
Training	284309	283360
Validation	65156	63158
Testing	73662	73308

Table 2: Number of sequences with embeddings in each class before and after filtering the data set

Class	Original	After filtering
0	216595	212129
1	212729	207697

4.3 Correlation analysis of embeddings' components

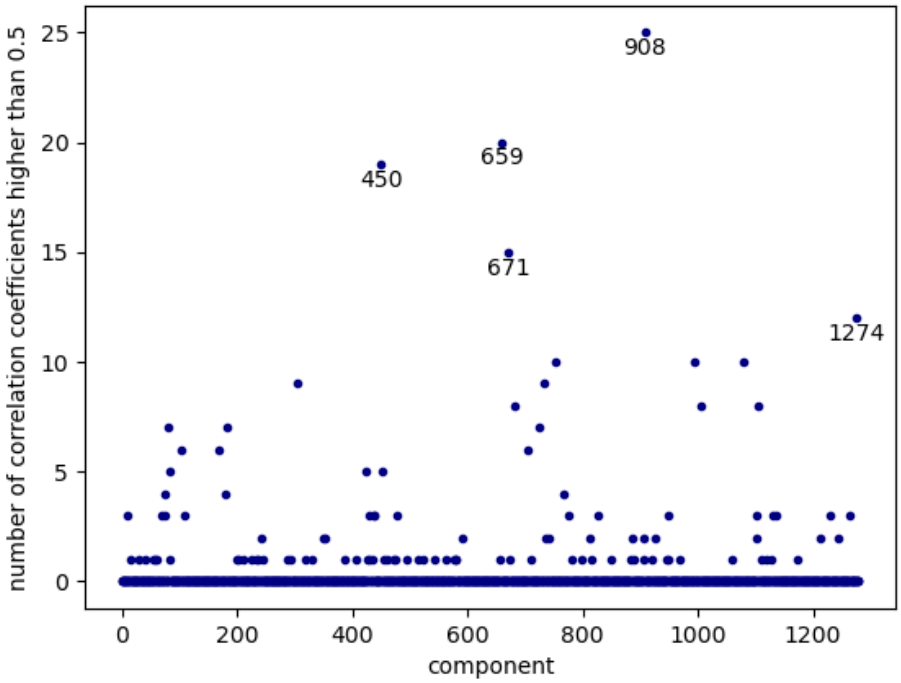

Figure 1: Plot of ESM-1b components that have got high correlation coefficients with ProtTrans components

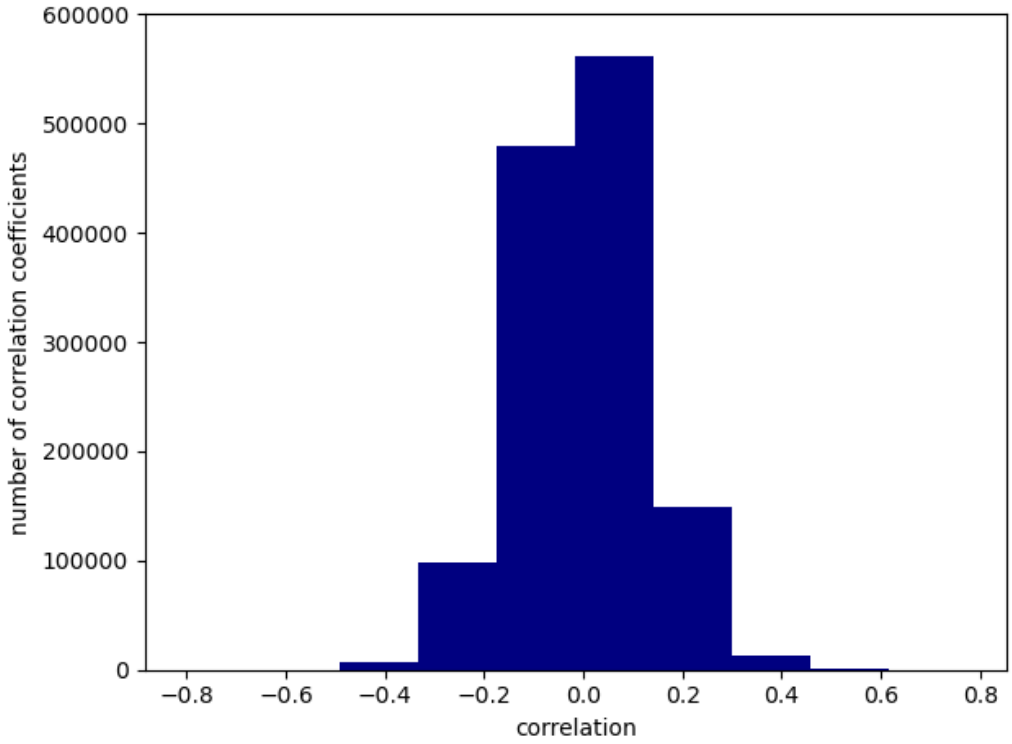

Figure 2: Histogram of correlation coefficients between ESM-1b and ProtTrans components

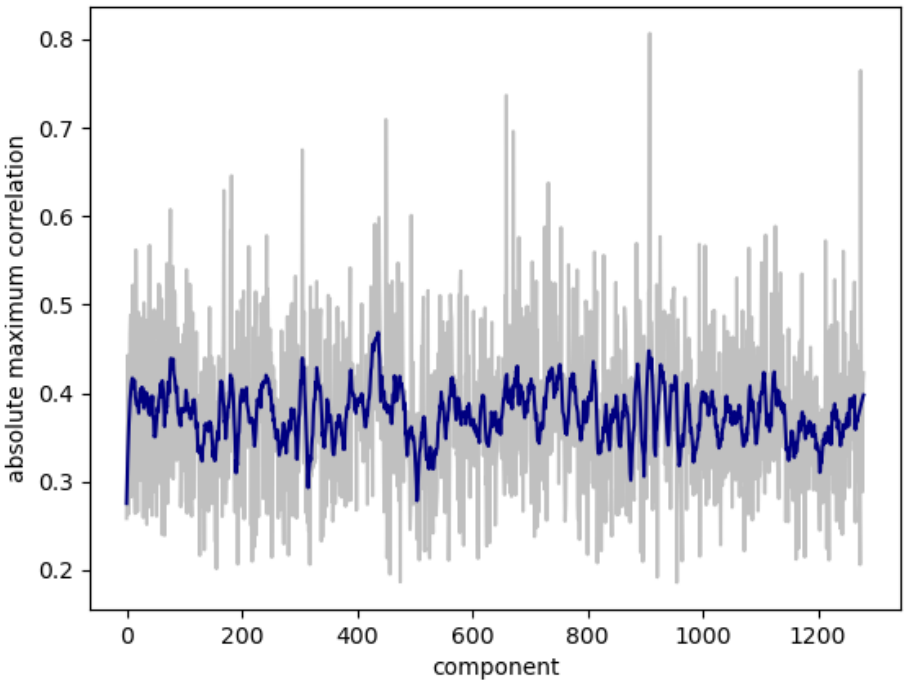

Figure 3: Plot of ESM-1b components' maximum correlation coefficients with ProtTrans components

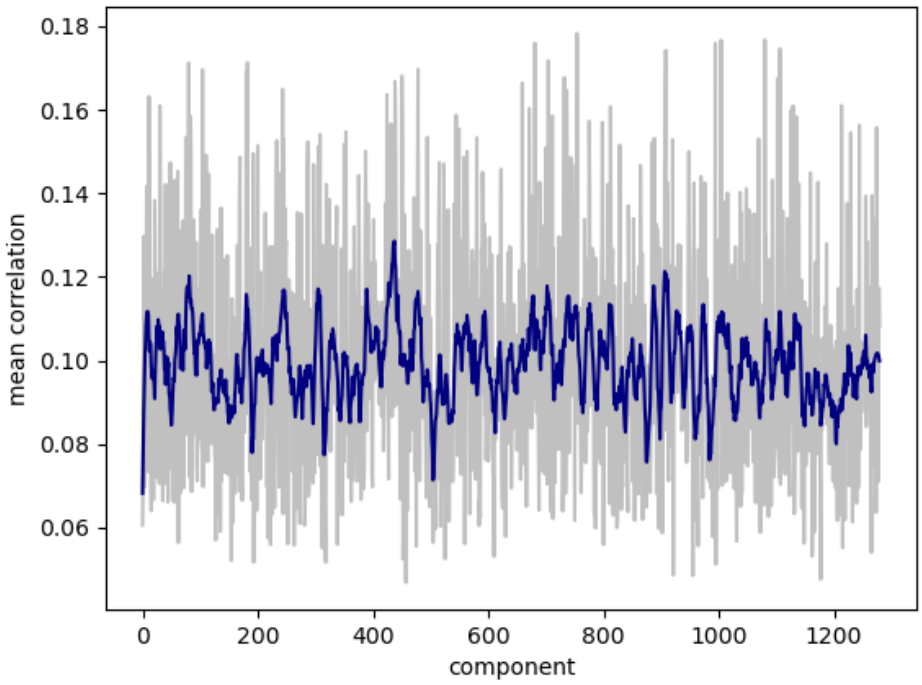

Figure 4: Plot of ESM-1b components' averaged correlation coefficients with ProtTrans components

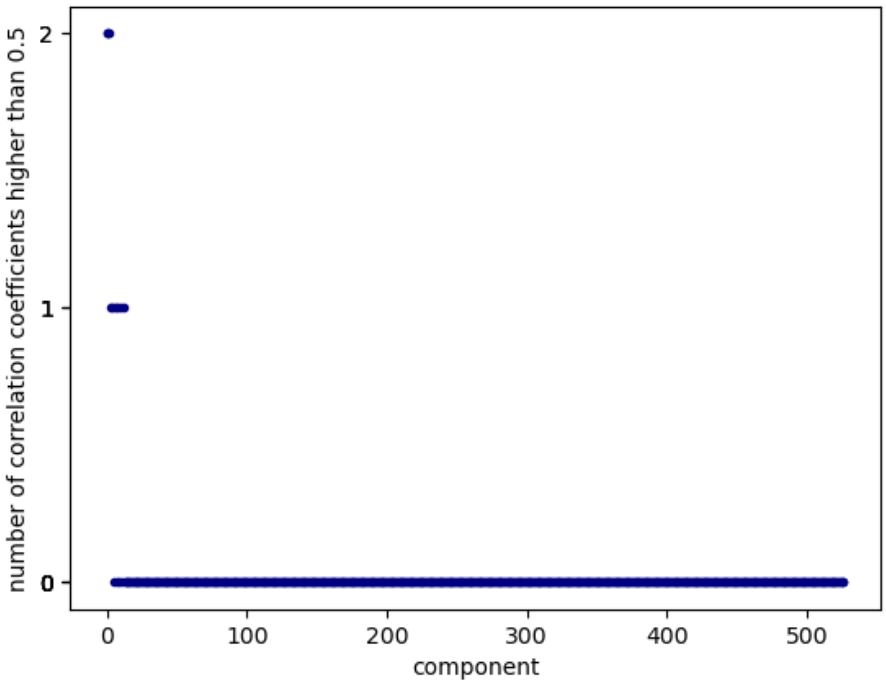

Figure 5: Plot of ESM-1b principal components (explaining 95% of data variation) that have got high correlation coefficients with ProtTrans principal components (95%)

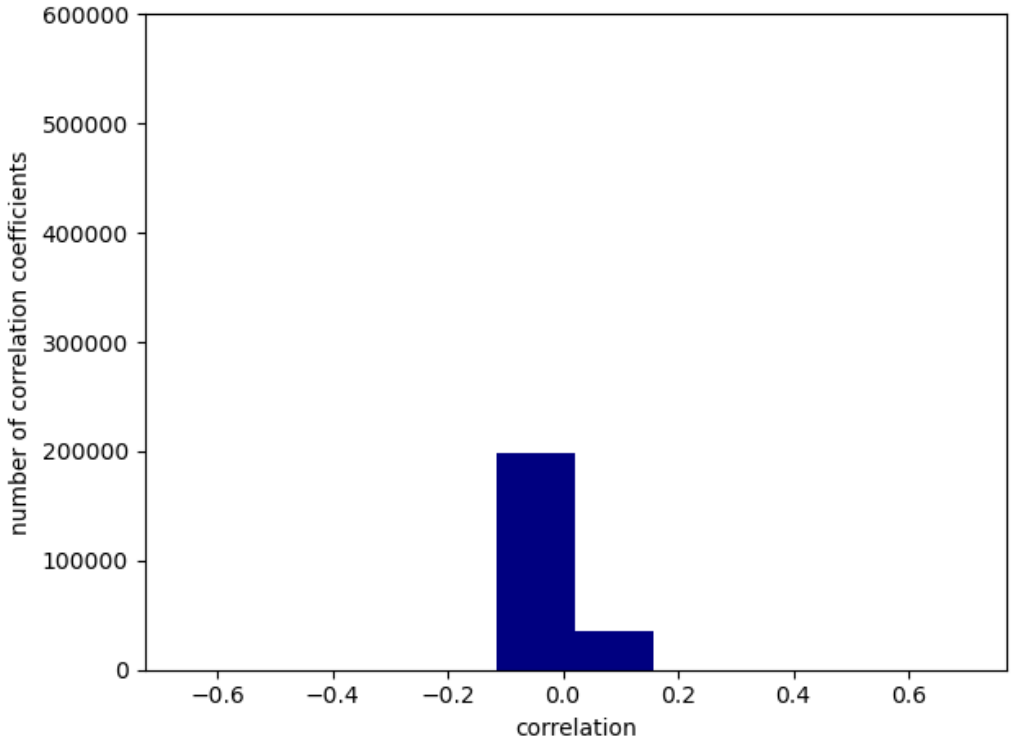

Figure 6: Histogram of correlation coefficients between ESM-1b and ProtTrans principal components (95%)

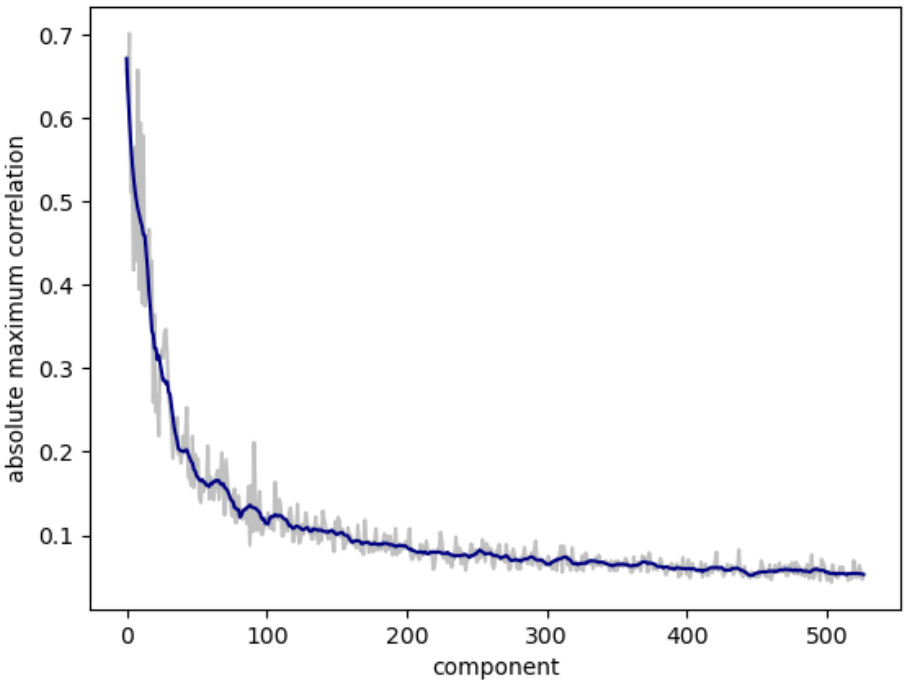

Figure 7: Plot of ESM-1b principal components' (95%) maximum correlation coefficients with ProtTrans principal components (95%)

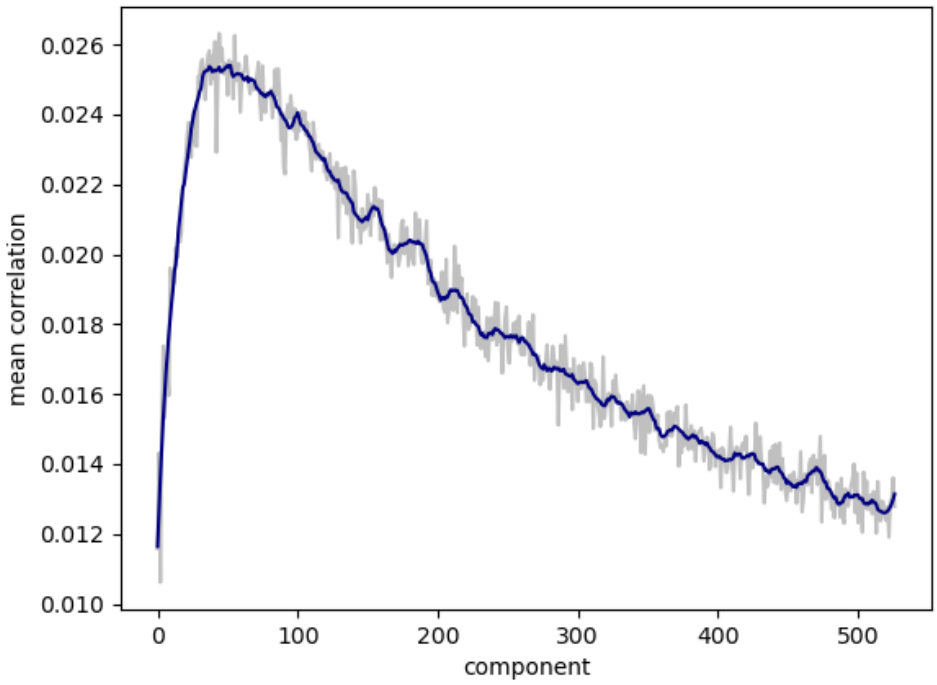

Figure 8: Plot of ESM-1b principal components' (95%) averaged correlation coefficients with ProtTrans principal components (95%)

4.4 Analysed representations

Firstly, principal components of ESM-1b and ProtTrans mean embeddings were retrieved and taken as input for the model. In particular, principal components that explained 95 percent and 100 percent of the data variance were taken. The data set for these representations analysis was picked the same as for the SLP analysis in the previous work.

Table 3: Sizes of the analysed principal components vectors

	ESM-1b	ProtTrans
95%	540	453
100%	1280	1024

Furthermore, both protein language models - ESM-1b and ProtTrans - provide per token or per residue representations - each amino acid of the protein gets 1280 or 1024-dimensional vector from ESM-1b or ProtTrans model respectively. Therefore, each protein is originally represented by $m \times n$ matrix, where m is the number of dimensions of the chosen type of embedding and n is the number of amino acids that compose the protein. These representations were processed to get vectors of the same dimension for each protein in the data set. The representations that were included in the analysis:

1. Mean ESM-1b and ProtTrans
2. Joined mean ESM-1b and ProtTrans
3. Normalised mean ESM-1b and ProtTrans
4. Joined normalised mean ESM-1b and ProtTrans
5. Median ESM-1b and ProtTrans
6. Minimum, median, and maximum ESM-1b and ProtTrans
7. Quantiles (including minimum and maximum) ESM-1b and ProtTrans
8. Quantiles (including minimum and maximum) and mean ESM-1b and ProtTrans
9. Octiles (including minimum and maximum) ESM-1b and ProtTrans

Table 4: Sizes of the analysed representations' vectors

Representation	ESM-1b	ProtTrans
Mean	1280	1024
Joined mean	2304	
Median	1280	1024
Minimum, median, maximum	3840	3072
Quantiles	6400	5120
Quantiles and mean	7680	6144
Octiles	11520	9216

4.5 Analysed architectures

Table 5: Models that were tested with ESM-1b embeddings input

Model	Number of hidden layers	Size of hidden layers
C2H2_h640-320	2	640, 320
C2H2_h320-160	2	320, 160
C2H1_h640	1	640
C2H1_h320	1	320
C2H1_h160	1	160
SLP_ESM-1b	0	-

Table 6: Models that were tested with ProtTrans embeddings input

Model	Number of hidden layers	Size of hidden layers
C2H2_h512-256	2	512, 256
C2H2_h256-128	2	256, 128
C2H1_h512	1	512
C2H1_h256	1	256
C2H1_h128	1	128
SLP_ProtTrans	0	-

5 Results

5.1 Representation analysis

Out of the scope of this representation and model architecture analysis for binary classification, there were several attempts made to overtrain the classification model to predict three thermostability classes (one more class added after dividing the zero-labelled class at the temperature threshold of 40 degrees Celsius). The purpose of overtraining was to check whether the selected architecture has potential to be trained for the multiclass classification problem optimally. The usage of principal components of protein embeddings showed that overfitting can be done successfully, thus this input was used further.

However, it was decided to check whether a vector of principal components could be a suitable input for the binary classification task. The testing stage metrics showed worse model's performance than using the original representations (Tables 7, 8 and 9).

Table 7: The comparison of scores between models trained with ESM-1b and ProtTrans mean representations of an unfiltered from duplicates data set

	ESM-1b	ProtTrans
MCC	0.843	0.902
Accuracy	0.922	0.951
Loss	0.208	0.128
Precision	0.919	0.949
Recall	0.921	0.951
ROC AUC	0.979	0.990

Table 8: The comparison of scores between models trained with ESM-1b and ProtTrans mean representations' principal components that account for 95% of the data set variance

	ESM-1b	ProtTrans
MCC	0.699	0.767
Accuracy	0.845	0.880
Loss	0.383	0.442
Precision	0.910	0.940
Recall	0.768	0.813
ROC AUC	0.901	0.945

By comparing SLP model's trained with ESM-1b embeddings results with results of SLP trained with ProtTrans, it was observed that the latter model performs better (Table 10).

Table 9: The comparison of scores between models trained with ESM-1b and ProtTrans mean representations' principal components that account for 100% of the data set variance

	ESM-1b	ProtTrans
MCC	0.698	0.766
Accuracy	0.845	0.879
Loss	0.382	0.443
Precision	0.909	0.939
Recall	0.767	0.812
ROC AUC	0.901	0.943

Table 10: The comparison of scores between models trained with ESM-1b and ProtTrans mean representations

	ESM-1b	ProtTrans
MCC	0.843	0.901
Accuracy	0.921	0.951
Loss	0.208	0.128
Precision	0.921	0.949
Recall	0.917	0.949
ROC AUC	0.979	0.990

Before joining the embeddings, the normalisation of ESM-1b and ProtTrans vectors was done. Normalised representations were taken as input to the model with the same SLP architecture. For both types of embeddings the results were improved (Tables 10 and 11).

Table 11: The comparison of SLPs', which were trained with normalised ESM-1b and ProtTrans mean representations, testing stage scores

	ESM-1b	ProtTrans
MCC	0.858	0.915
Accuracy	0.929	0.957
Loss	0.248	0.143
Precision	0.923	0.951
Recall	0.931	0.962
ROC AUC	0.982	0.991

After joining ESM-1b and ProtTrans mean embeddings, an SLP was trained using these joined representations. The results of this model were similar to the model's trained using only ProtTrans embeddings, yet the results were not improved (Tables 10 and 12).

However, joining the normalised ESM-1b and ProtTrans mean representations showed the best results. Since joined representations require generation of ESM-1b embeddings, this

type of representation does not solve the length limitation problem. Nevertheless slightly improved results can be observed when normalised representations are used, the process of normalisation depends on the data set, which is not convenient in the process of development until the final data set is established. Therefore, the optimal choice for this stage of development was ProtTrans embeddings.

Table 12: The comparison of testing stage scores between models trained with ESM-1b and ProtTrans mean representations

	Joined	Normalised joined
MCC	0.899	0.920
Accuracy	0.949	0.960
Loss	0.131	0.139
Precision	0.945	0.954
Recall	0.951	0.964
ROC AUC	0.991	0.992

Nonetheless, ProtTrans already demonstrated the impact for the model's improvement of performance, it was decided to finish up the different representation and architecture analysis using both types of embeddings. Although, the results of the consequent analysis did not change the conclusion regarding ProtTrans influence for the results for any variation of analysed representations (listed in the section 4.4) - in all cases model that used ProtTrans embeddings performed significantly better (Figures 9 and 10).

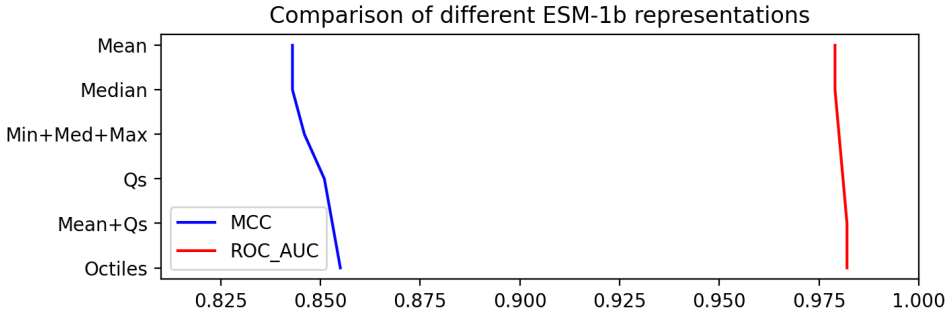

Figure 9: Comparison of SLP models', which were trained with different ESM-1b representations, MCC and ROC AUC scores

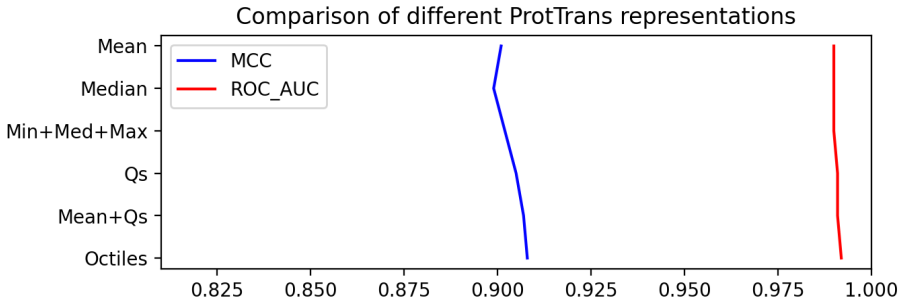

Figure 10: Comparison of SLP models', which were trained with different ProtTrans representations, MCC and ROC AUC scores

5.2 Architecture analysis

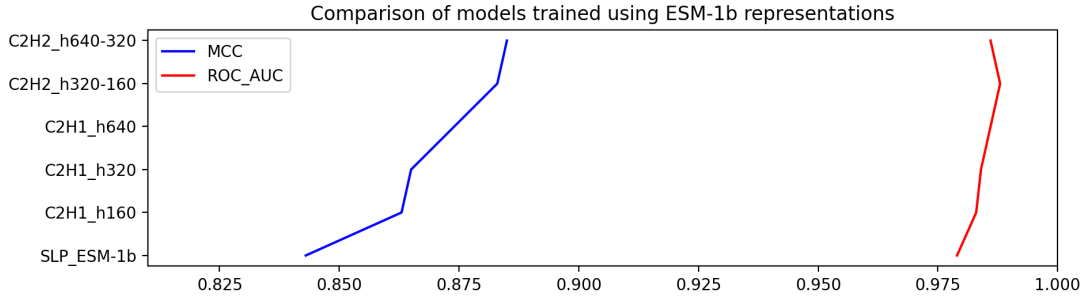

Figure 11: Comparison of models', which were trained using ESM-1b embeddings, MCC and ROC AUC scores

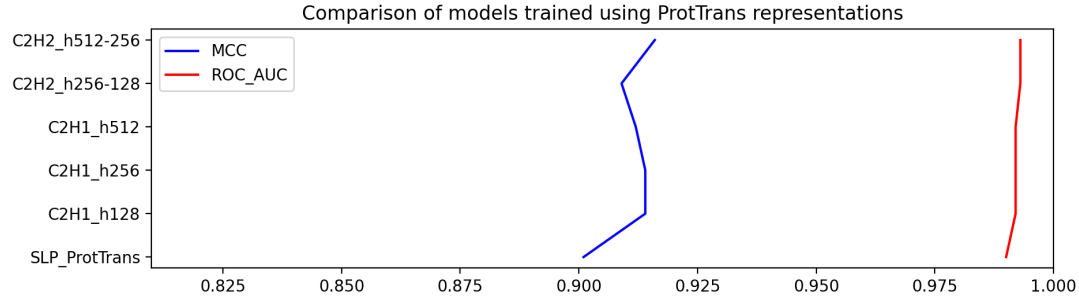

Figure 12: Comparison of models', which were trained using ProtTrans embeddings, MCC and ROC AUC scores

6 Conclusions

The results provided following conclusions:

1.

7 Availability

The code that was used to receive the results of this work can be found in the designated Github repository: https://github.com/ievapudz/Course_Work_Project.

References