Deep Learning - Lab06

INSTITUTE OF MOLECULAR BIOLOGY

Howard Hughes Medical Institute

ZIMIN FOUNDATION

SMTB

INTRODUCTION

ESM (Evolutionary Scale Modeling) - это маскировочная языковая модель, принимающая на вход белковые последовательности и выдающая эмбедингы - числа, характеризующие последовательность [1]. Эмбеддинги используются для предсказания пропущенной аминокислоты, флуоресценции, стабильности и других целей.
В данной работе мы занимались изучением свойств ESM. Нас также интересовало, на какие параметры нейросеть обращает больше внимания при анализе белковой последовательности. С каждого слоя ESM мы получили эмбеддинги, которые использовали для предсказания стабильности [2] и флюоресцентности [3] белка с помощью своей однослойной нейросети, подобно авторам статьи по ESM. Мы выяснили, что эмбеддинги последнего слоя не являются наиболее оптимальными для этих целей [1].
Также мы посмотрели на эмбеддинги базы данных с белками из разных классов и сравнили с эмбеддингами этих же белков, но с перемешанной последовательностью.

ESM (Evolutionary Scale Modeling) is a masked language model. It receives protein sequences as input and computes embeddings as output. Embeddings are numerical representations, characterizing the sequence [1]. These embeddings are used for predicting missing amino acids, fluorescence, stability, and other properties.
In this study, we focused on exploring the properties of ESM. We were also interested in identifying which embedding positions the neural network pays more attention to when analyzing protein sequences. From each layer of ESM, we obtained embeddings, which were used for predicting protein stability [2] and fluorescence [3] with a single-layer neural network, similar to the authors of the ESM paper. We found that embeddings from the final layer are not always the best performing ones [1].
Additionally, we looked at the embeddings of a set of proteins from different classes and compared them to the embeddings of the same proteins but with a shuffled sequence.

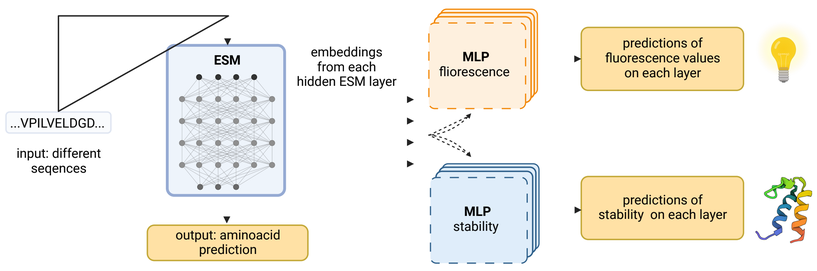
GOAL AND OBJECTIVES

- Проверить гипотезу: обращает ли ESM наибольшее внимание на пиковые значения эмбеддингов;
- Выяснить, какой слой ESM оптимален для предсказания:
  - Флуоресценции белка;
  - Стабильности белка;
- Установить, как меняются значения эмбеддингов в зависимости от слоя;

- Test the hypothesis: Does ESM pay the most attention to the peak values of embeddings;
- Find out which ESM layer is optimal for prediction:
  - Protein fluorescence;
  - protein stability;
- Set how embedding values change depending on the layer;

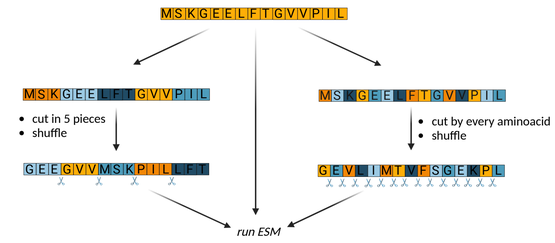
METHODS

Fluorescence and stability predictions



Для установления слоя ESM с наиболее точными предсказаниями флуоресценции и стабильности белка мы использовали значения эмбеддингов из каждого скрытого слоя ESM, а также дополнительные однослойные нейросети, которые мы предварительно обучили определять флуоресценцию и стабильность белка по эмбеддингам. Датасет из набора данных TAPE-dataset
In order to find the ESM layer with the most accurate predictions of the fluorescence and the stability of a protein, we used embeddings from each layer of ESM as input to a single-layer neural networks, trained to predict protein fluorescence and stability of a protein. The datasets were taken from the TAPE dataset.

Homology



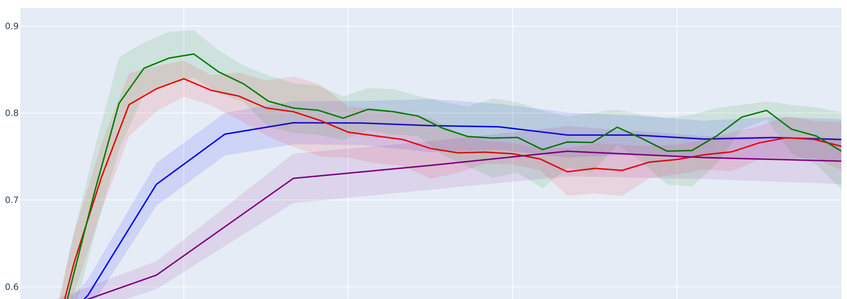
Мы изменили исходную последовательность белка двумя способами и сравнили эмбеддинги.
We changed the original protein sequence in two ways and compared the embeddings.

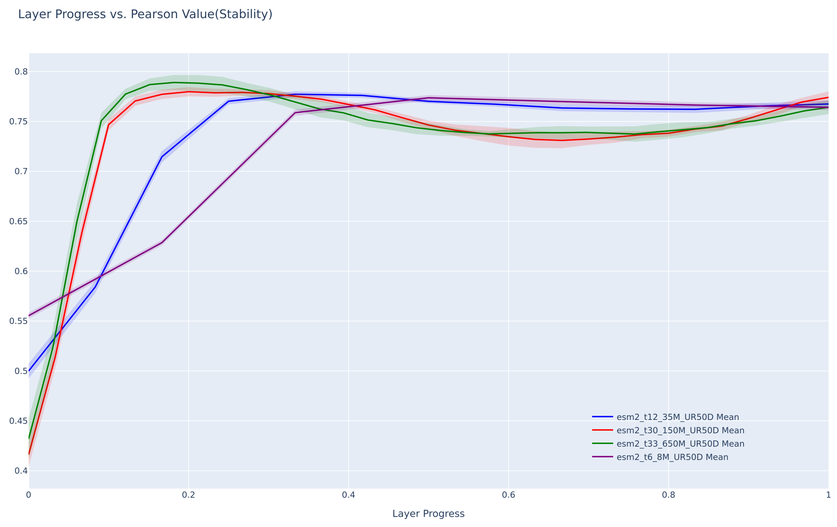
RESULTS

Results for stability and fluorescence

Из полученных графиков мы заметили, что значение критерия Пирсона на 6-8 слоях выше, чем значение на 33-ем слое, соответственно предсказание флуоресценции и стабильности на ранних слоях возможно с лучшей точностью, чем на последних.

From the obtained graphs, we noticed that the Pearson correlation coefficient is higher on layers 6-8 compared to layer 33. Therefore, predicting fluorescence and stability on earlier layers may be more accurate than on the later ones.

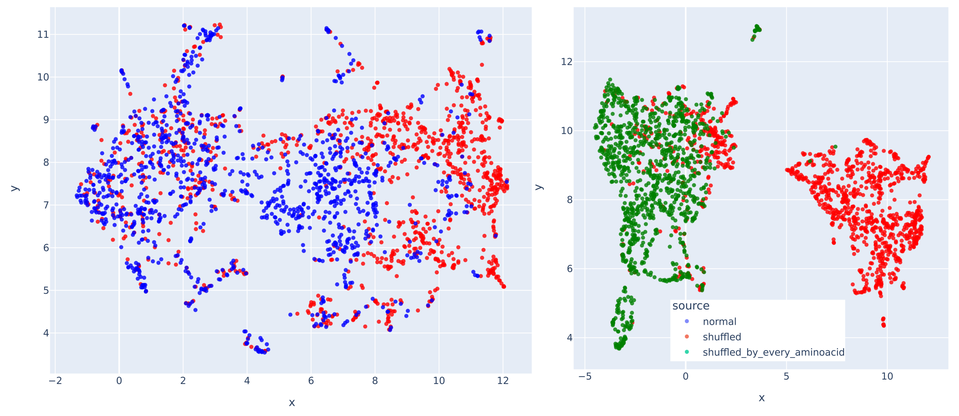




Results for homology

На графиках можно увидеть, что кластеризация белков через значение эмбеддингов ESM зависит от того, каким способом мы изменили изначальную последовательность. Следовательно, ESM способна заметить разницу между существующим белком и искусственно созданной последовательностью. Так мы дополнительно хотели показать, что для ESM важна не только встречаемость аминокислот, но и их порядок в последовательности.

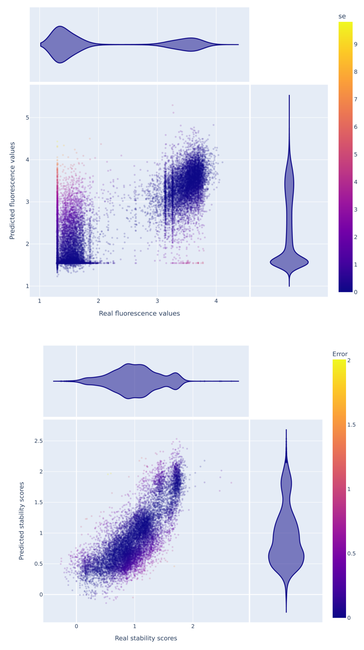
In the graphs, it can be observed that the clustering of proteins by ESM embedding depends on modification of the original sequence. Consequently, ESM is capable of detecting the difference between an existing protein and an artificially generated sequence. Thus, we aimed to demonstrate that for ESM, both the occurrence and the order of amino acids in the sequence are important.

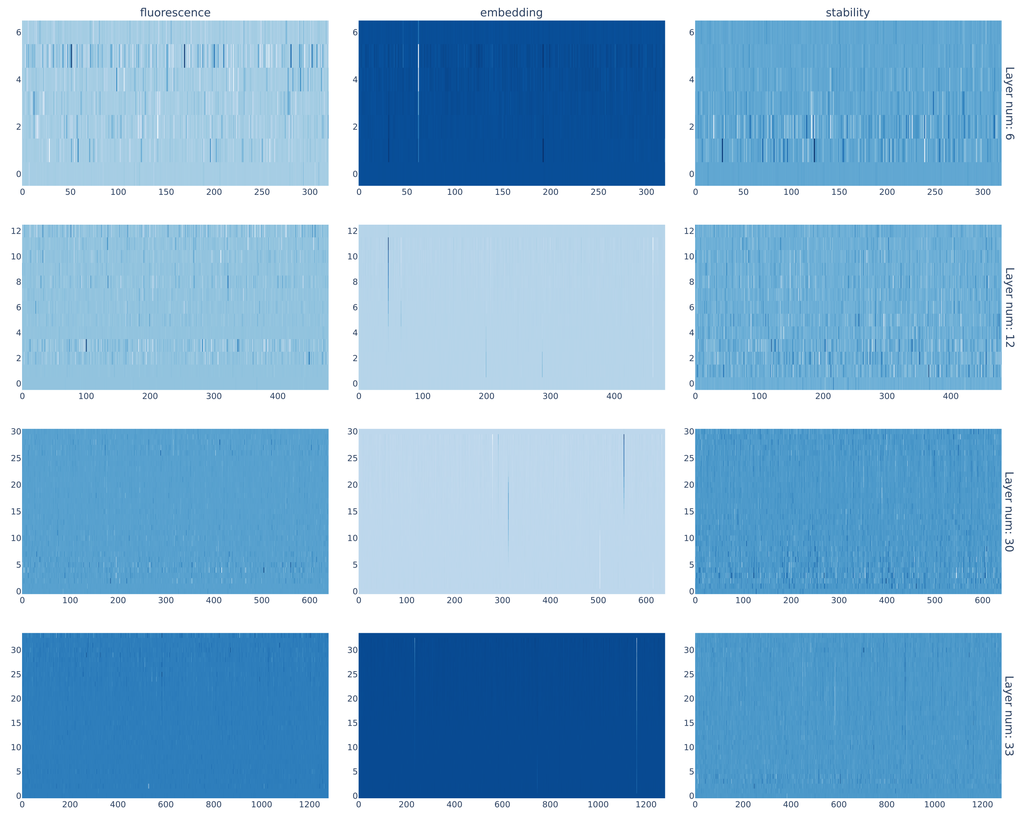


Explainabiliy

На графиках ниже показано сравнение фактических значений флуоресценции и стабильности и значений, предсказанных моделью, обученной на слое 8 ESM2-t33.

The graphs below depict a comparison between actual fluorescence and stability values and those predicted by the model trained on layer 8 of ESM2-t33.





Для объяснения полученных результатов мы использовали метод DeepLift и сравнивали полученные значения эмбеддингов с референсом - “baseline”, который является пустой строкой (последовательностью с замаскированными значениями).В первом и третьем столбце теум цвтом указано насколько ESM обращает внимание на ту или иную позицию в каждом слое модели, во втором столбце показаны значения эмбеддингов.

To explain the obtained results, we uses DeepLift and compared the obtained embedding values with a reference - the "baseline", which is an empty string (a sequence with masked values).The first and third columns indicate in a dark shade how much attention ESM pays to each position in each layer of the model, while the second column displays the embedding values.

REFERENCES

1. Rives, Alexander, et al. "Biological Structure and Function Emerge from Scaling Unsupervised Learning to 250 Million Protein Sequences." *PNAS*, 2019. DOI: 10.1101/622803.
2. Rocklin, Gabriel J. et al. "Global Analysis of Protein Folding Using Massively Parallel Design, Synthesis, and Testing." *Science*, vol. 357, no. 6347, 2017, pp. 168-175.
3. Sarkisyan, Karen S. et al. "Local Fitness Landscape of the Green Fluorescent Protein." *Nature*, vol. 533, no. 7603, 2016, p. 397.

ACKNOWLEDGMENT

Thanks for:

- Ilya and Roman for their great patience, care, skills, explanation, jokes and cool music
- Olga Kalinina for funding them
- Sofas on the 2nd floor in the Institute
- Alexander Kharkhota for his moral support

CONCLUSIONS

- Последний слой ESM не является самым лучшим для всех целей;
- 7 - 8 слои являются оптимальными для предсказания флуоресцентности и стабильности белков;
- На первых слоях ESM обращает внимание только на аминокислотный состав последовательности. Чем больше номер слоя, тем больше значения придается белковой структуре;

- Last ESM layer is not always the best layer;
- 7th - 8th layer are optimal for predicting the fluorescence and stability of proteins;
- In the first layers, ESM pays attention only to the amino acid composition of the sequence. The higher the layer number, the more importance is attached to the protein structure;

SCAN ME