StructPLM: Enhancing Protein Representations with Structural Information

Daria Frolova
Marina Pak
Ivan Oseledets
Dmitry Ivankov

Protein representations in bioinformatics

Growth of the number of available **protein sequences** (primary structure)

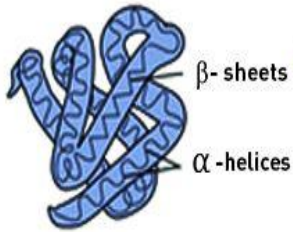
Transformers (ESM2 [2], ProtT5 [3])

+ deep networks
- no use of 3D structure

Protein **structural information** has become available with AlphaFold2 [1] (tertiary structure)

Graphs

+ use of 3D structure
- need for equivariance
- shallow networks



Our idea: add protein structural information to a transformer model

Image credits: https://byjus.com/biology/proteins-structure-and-functions/

Existing attempts to use 3D structure in representations

- ProstT5 [4]
  - 3Di alphabet (1D-strings representing protein 3D structure (Foldseek [5])
  - Train to translate between 3Di and amino acid sequences
  - Mostly performs worse then sequence-only ProtT5 model on downstream tasks

- S-PLM [6]
  - Represents structure as C_α contact maps, apply CNN to it
  - Aligns sequence and structure embeddings with contrastive learning
  - Extensive evaluation on various downstream tasks

Proposed StructPLM

- Amino acid side chains can exist only in a few positions - rotamers
- Add new tokens
  - amino acid type
  - backbone torsion angles φ, ψ
  - side-chain rotamer type
  - side-chain nonrotameric angle (if any)
- Binarize angles into 2 degree bins
- Compute smoothed cross-entropy loss on angle tokens to perform angle regression:
  - Penalize angle predictions based on the angle difference
  - It is better to predict 122° than 160° (when we have angle 120°)

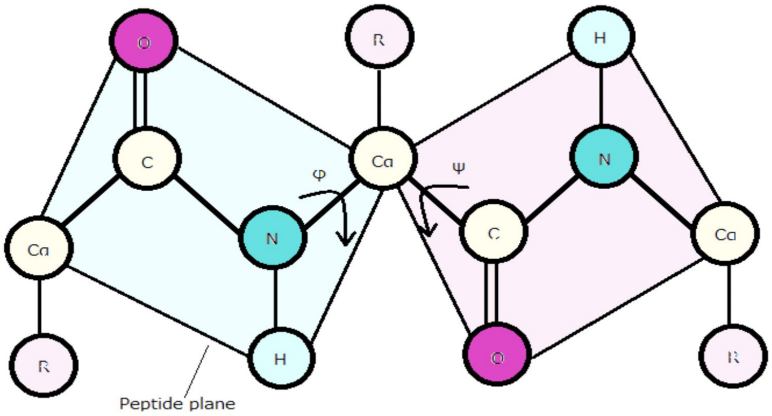


Image credits: http://cib.cf.ocha.ac.jp/bitool/DIHED2/

Experimental setup

Datasets: **AlphaFold database**: >500k structures, generated with AlphaFold2 - used for pre-training

Protein Data Bank: ~41k experimental structures - used for finetuning

Model: 12-layer (87 M parameters) RoBERTa model

Downstream task:

Per-residue prediction: prediction of protein stability change (ΔΔG) due to single mutations

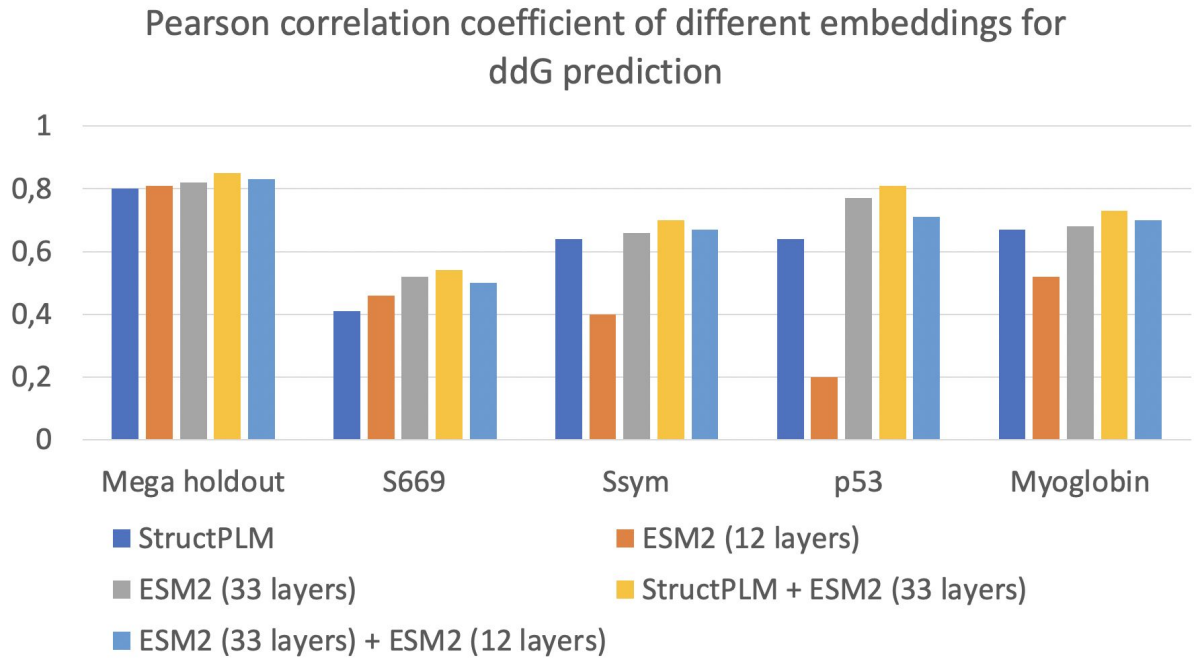
We follow the setup of ABYSSAL [7], a top-performing neural network for ΔΔG prediction. It works upon ESM2 embeddings.

Train data: Mega dataset [8]
Test data: Mega dataset holdout, S669 [9], ssym [10], p53 [11], Myoglobin [12]

Our approach:

1. Train ABYSSAL on **StructPLM** embeddings
2. Train ABYSSAL on **concatenation of ESM2 and StructPLM** embeddings

Experimental results



StructPLM is mostly better than a small ESM2 model (12 layers)

The concatenation of StructPLM and ESM2 embeddings increases the performance on a downstream task of ΔΔG prediction

Conclusion

- We proposed **StructPLM** that uses structural information inside pLM
- Small StructPLM model produces high-quality embeddings
- Even embeddings from a small StructPLM model can **enhance ESM2 embeddings** on a downstream task

Further research:

- Perform extensive experiments with various downstream tasks
- Compare to existing structural models S-PLM and ProstT5
- Consider increasing the size of the model for better performance

References

[1] Jumper, J., et. al. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589.

[2] Rives, A., et. al. 2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118.

[3] Elnaggar, A., et. al. (2021). Prottrans: Toward understanding the language of life through self-supervised learning. IEEE transactions on pattern analysis and machine intelligence, 44(10), 7112-7127.

[4] Heinzinger, M., et. al. (2023). ProstT5: Bilingual Language Model for Protein Sequence and Structure. bioRxiv, 2023-07.

[5] van Kempen, M., et. al. (2023). Fast and accurate protein structure search with Foldseek. Nature Biotechnology, 1-4.

[6] Wang, D., et. al. (2023). S-PLM: Structure-aware Protein Language Model via Contrastive Learning between Sequence and Structure. bioRxiv, 2023-08.

[7] Pak, M. A., et. al. (2023). The new mega dataset combined with a deep neural network makes progress in predicting the impact of single mutations on protein stability. bioRxiv, 2022-12.

[8] Tsuboyama, K., et. al. (2023). Mega-scale experimental analysis of protein folding stability in biology and design. Nature, 1-11.

[9] Pancotti, C., et. al. (2022). Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. Briefings in Bioinformatics, 23(2), bbab555.

[10] Pires, D. E., et. al. (2014). mCSM: predicting the effects of mutations in proteins using graph-based signatures. Bioinformatics, 30(3), 335-342.

[11] Pucci, F., et. al. (2018). Quantification of biases in predictions of protein stability changes upon mutations. Bioinformatics, 34(21), 3659-3665.

[12] Kepp, K. P. (2015). Towards a “Golden Standard” for computing globin stability: Stability and structure sensitivity of myoglobin mutants. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics, 1854(10), 1239-1248.

Contacts

Daria Frolova

Researcher at Ligand Pro

PhD student at Skolkovo Institute of Science and Technology

Daria.Frolova@skoltech.ru