A Comprehensive Review of Protein Language Models

- **Authors of the paper:** Lei Wang, Xudong Li, Han Zhang, Jinyi Wang, Dingkang Jiang, Zhidong Xue and Yan Wang
- **Affiliation:** Huazhong University of Science and Technology

- **Journal:** Arxiv preprint
- **Published:** February 2025

8.12.2025

presented by Magdaléna Turinská

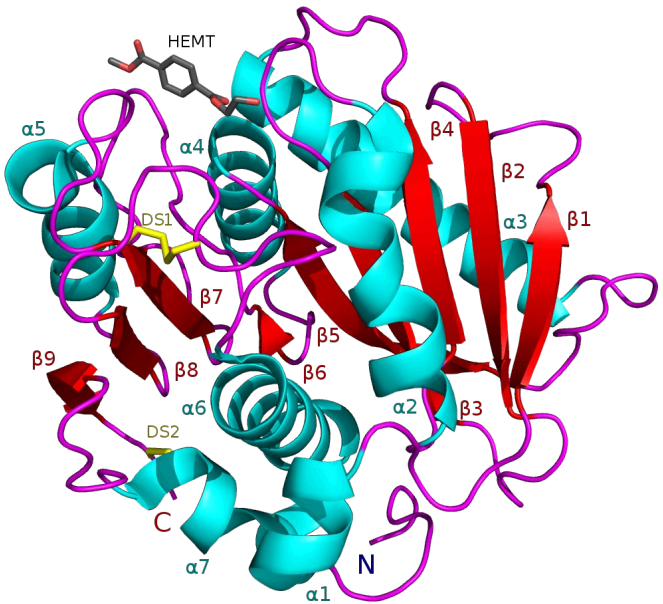
Table of contents

Motivation

- Designing novel proteins can solve the **biomedical and environmental** problems

- Biological experimental research methods are very time-consuming and expensive opposite to computational biology.

PETase

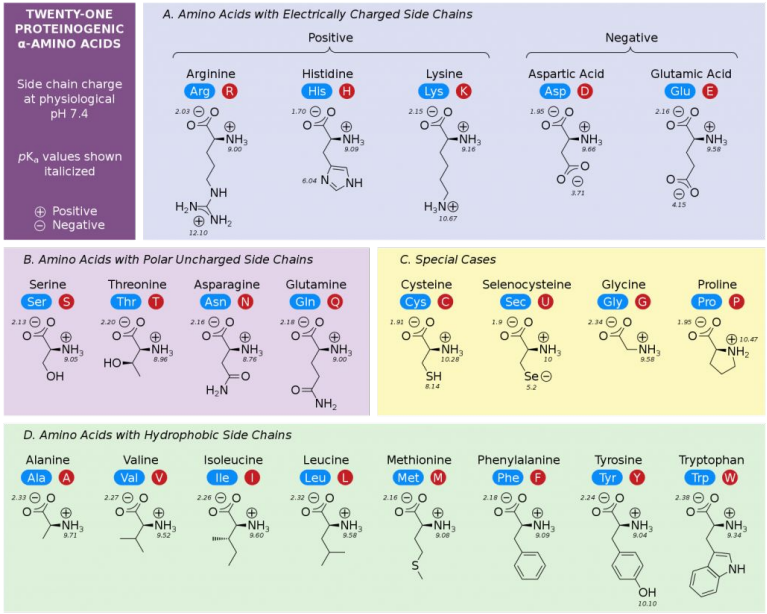


By Keministi - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=68401651

Sequence of protein

- Protein is linear chain of Amino-acids
- Only 20 different amino-acids
- We use edit distance to define similarity.
- $20^{300} \approx 10^{390} > 10^{80}$
- Multiple sequence alignment (MSA)
  - O(Length^{Nseqs}) - without heuristics

Amino acids table

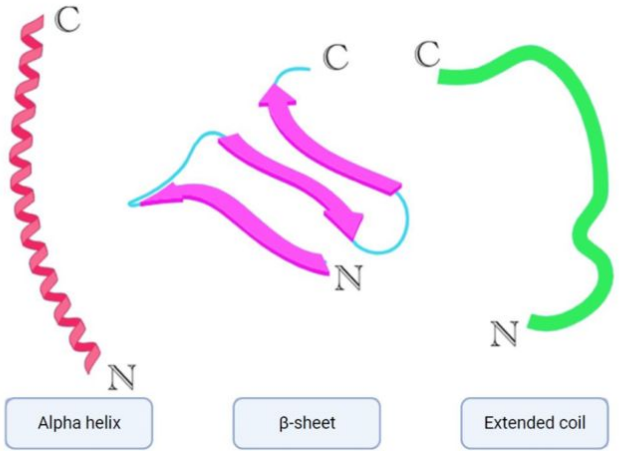


https://chemistrytalk.org/amino-acid-chart/

Structure of protein

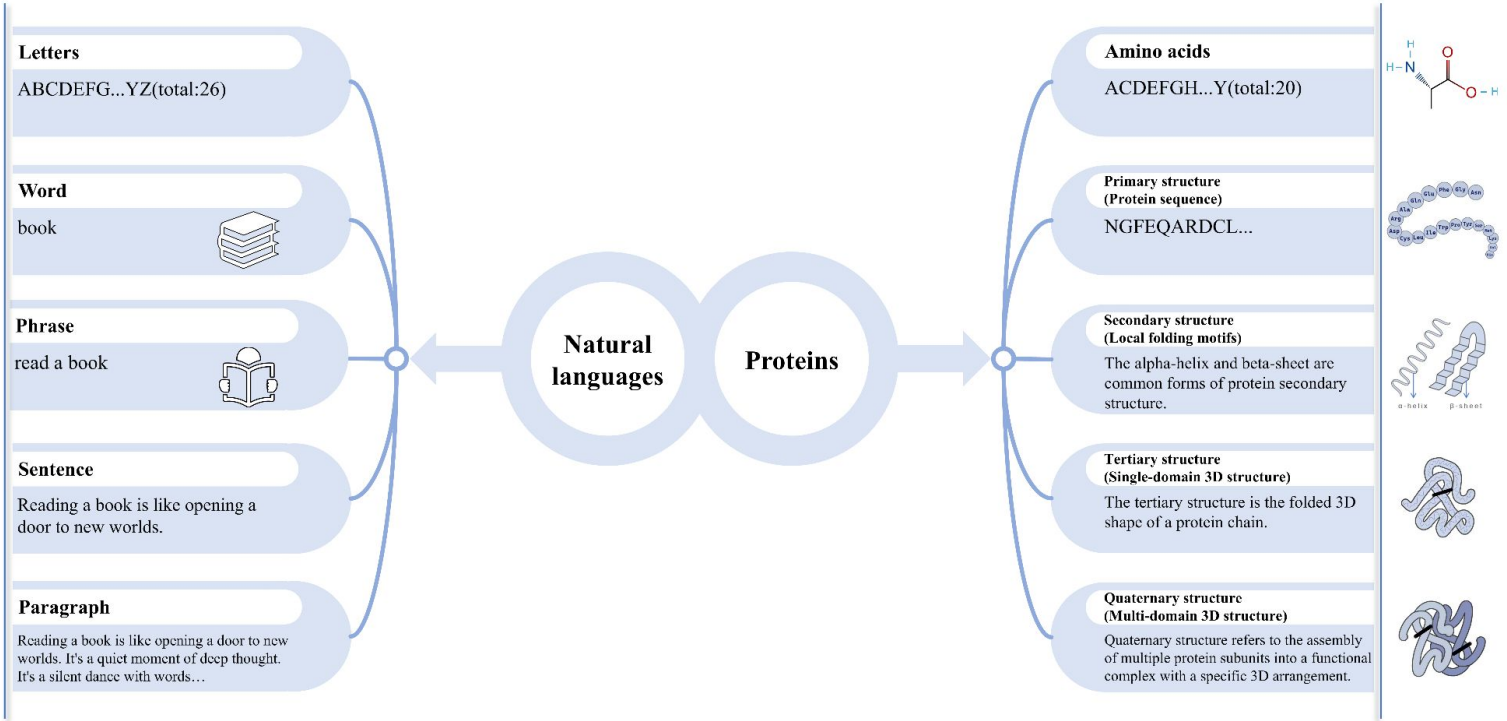
- Structure (folded protein) is spontaneously formed from sequence
- (Stable) structured vs disorder protein
- Secondary structure of protein:
  - Alpha-helix,
  - Beta-sheet,
  - Coil
- Prediction of protein structure is hard problem - deepMind research

Secondary structures:



https://www.researchgate.net/figure/Simplified-models-of-alpha-helix-b-sheet-and-extended-coil-structures-Created-in_fig1_378710645

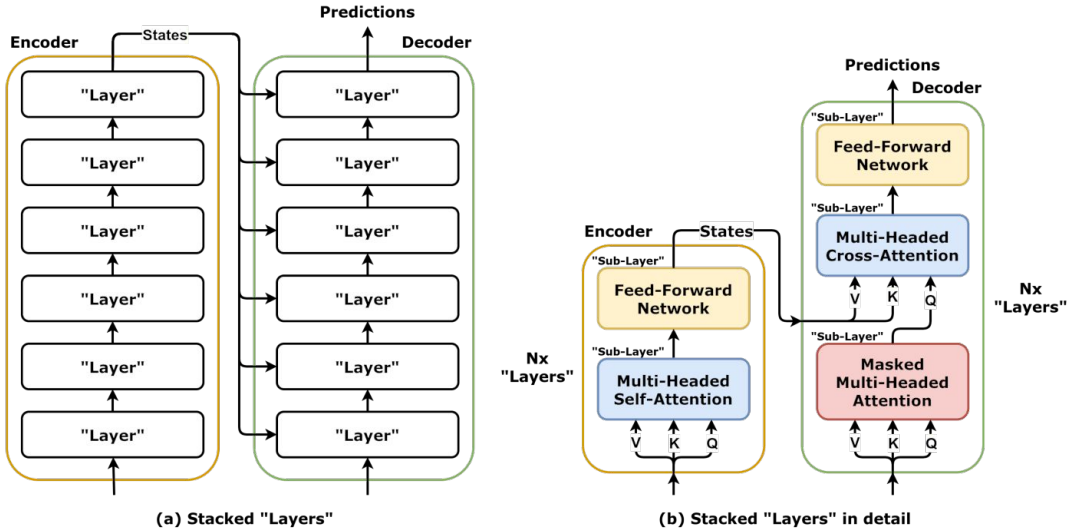
Conceptual similarities in natural languages and proteins



!!! Protein sequence is 1 string without whitespace/separators.

Transformer

- Artificial neural network
- Multi-head attention mechanism
- Encoder + Decoder



By dvgodoy - https://github.com/dvgodoy/dl-visuals/?tab=readme-ov-file, CC BY 4.0, https://commons.wikimedia.org/w/index.php?curid=151216018

Non-transformer-based Models

- Limitations:
  - Only handle fixed-length sequences
  - Typically require a large amount of labeled data

Model	Pretraining Dataset	Base Model	Params	Time	Code
CARP	UniRef50	CNN	600K-640M	2024.02	✓
MIF-ST	CATH	GNN	3.4M	2023.03	✓
ProSE	UniRef90, SCOP	LSTM	-	2021.06	✓
Seq2vec	-	CNN-LSTM	-	2020.09	✗
UDSMProt	UniProtKB/Swiss-Prot	AWD-LSTM	-	2020.01	✗
SeqVec	UniRef50	ELMo	-	2019.12	✓
UniRep	UniRef50	mLSTM	-	2019.10	✓
ProtVecX	UniRef50, UniProtKB/Swiss-Prot	ProVec	-	2019.03	✗
ProtVec	UniProtKB/Swiss-Prot	Skip-gram	-	2015.11	✗

TABLE I: Non-transformer-based models

Transformer-based Models

- Encoder-only (BERT - like)
  - Encode protein sequences into fixed-length vector representations
- Decoder-only (GPT - like)
  - Used for protein generation tasks
  - Auto-regressive models
  - ProptGPT2
- Encoder–Decoder (Text-to-Text Transfer Transformer - like)
  - Typically used for sequence-to-sequence tasks
  - Combined sequence and structure information

Encoder–Decoder: Positional encoding of amino-acids

- Absolute encoding
  - Simplicity
  - Limiting the model’s ability to handle sequences of varying lengths
    - Positions of amino-acids
    - Rotation matrix and precise distances
- Relative encoding
  - Relationship between 2 tokens

Models scaling

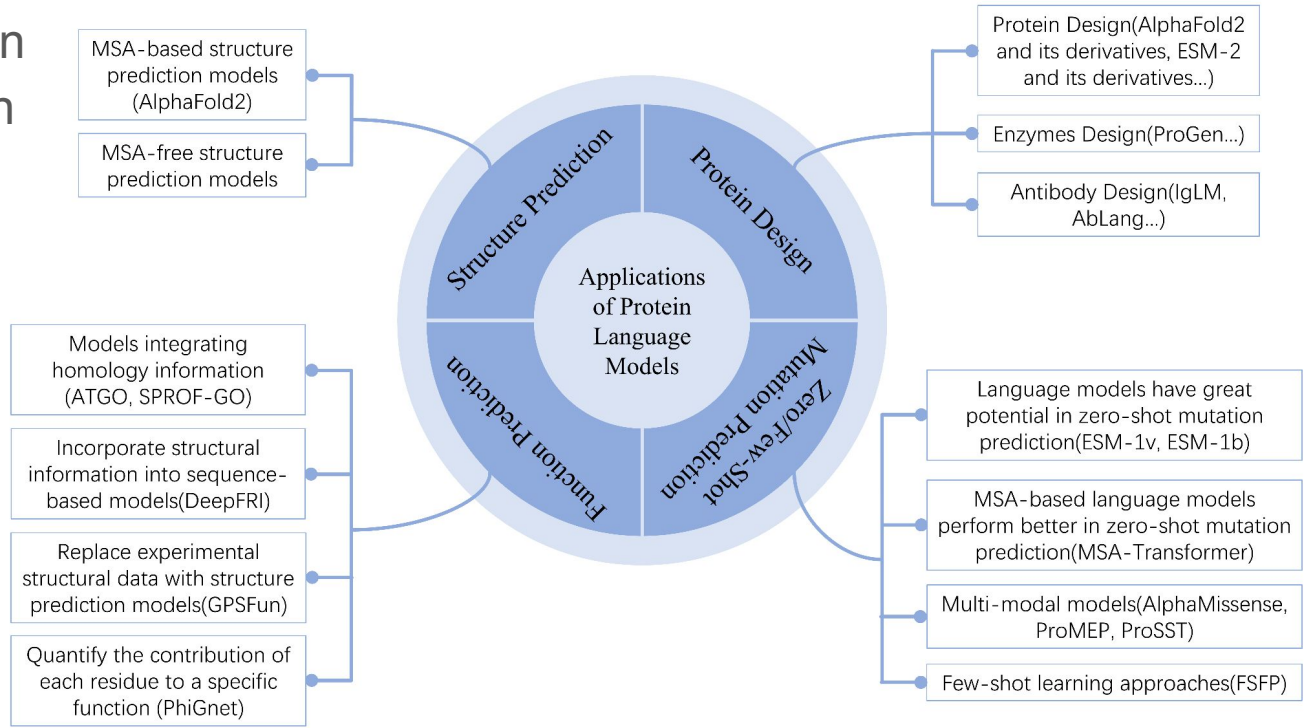
- Scaling PLMs brings bigger gains than scaling NLP models.
- PLMs are more prone to underfitting
- Further scaling of models can improve the performance of PLMs

Application of PLM

- Structure Prediction
- Function Prediction
- Protein Design

Future develop

Multimodal Models



ProptGPT2 - example model

ProtGPT2 is a deep unsupervised language model for protein design

- **Authors of the paper:** Noelia Ferruz, Steffen Schmidt & Birte Höcker
- **Affiliation:** Department of Biochemistry, University of Bayreuth, Bayreuth, Germany
- **Journal:** Nature Communications volume
- **Published:** July 2022

ProptGPT2

- Used to:
  - De novo protein design
  - engineering
- Model description
  - 36 layers
  - model dimensionality 1289
  - 738 milion parameters
  - trained on ~50 million non-redundant protein sequences
- Generates sequences in seconds
- Freel available: https://huggingface.co/nferruz/ProtGPT2

Example of prediction

- Used as HuggingFace transformer python package
- Best sequence = lower perplexity metrix
- Validate of new sequence by predict the structure

```
>>> from transformers import pipeline
>>> protgpt2 = pipeline('text-generation', model="nferruz/ProtGPT2")
# length is expressed in tokens, where each token has an average length of 4 amino acids.
>>> sequences = protgpt2("<|endoftext|>", max_length=100, do_sample=True, top_k=950,
                repetition_penalty=1.2, num_return_sequences=10, eos_token_id=0)
>>> for seq in sequences:
        print(seq):
{'generated_text': 'MINDLLDISRIISGKMTLDRAEVNLTAIARQVVEEQRQAAEAKSIQLLCSTPDTNHYVFG\nDFDRLKQTLWNLLSNAVKFTPSGGTVELELGYNAEGMEV-
{'generated_text': 'MQGDSSISSSNRMFT\nLCKPLTVANETSTLSTTRNSKSNKRVSKQRVNLAESPERNAPSPASIKTNETEEFSTIKT\nTNNEVLGYEPNYVSYDFVPMEK-
{'generated_text': 'M\nSTHVSLENTLASLQATFFSLEARHTALETQLLSTRTELAATKQELVRVQAEISRADAQAQ\nDLKAQILTLKEKADQAEVEAAAATQRAEESQAALEA-
{'generated_text': 'MVLLSTGPLPILFLGPSLAELNQKYQVVSDTLLRFTNTV\nTFNTLKFLGSDS\n'}
{'generated_text': 'M\nNNDEQPFIMSTSGYAGNTTSSMNSTSDFNTNNKSNTWSNRFSNFIAYFSGVGWFIGAISV\nIFFIIYVIVFLSRKTKPSGQKQYSRTERNNRDVDSI-
{'generated_text': 'M\nEAVYSFTITETGTGTVEVTPLDRTISGADIVYPPDTACVPLTVQPVINANGTWTLGSGCT\nGHFSVDTTGHVNCLTGGFGAAGVHTVIYTVETPYSG-
{'generated_text': 'M\nGLTTSGGARGFCSLAVLQELVPRPELLFVIDRAFHSGKHAVDMQVVDQEGLGDGVATLLY\nAHQGLYTCLLQAEARLLGREWAAVPALEPNFMESPL-
{'generated_text': 'M\nGAAGYTGSLILAALKQNPDIAVYALNRNDEKLKDVCGQYSNLKGQVCDLSNESQVEALLS\nGPRKTVVNLVGPYSFYGSRVLNACIEANCHYIDLTG-
{'generated_text': 'M\nKFPSLLLDSYLLVFFIFCSLGLYFSPKEFLSKSYTLLTFFGSLLFIVLVAFPYQSAISAS\nKYYYFPFPIQFFDIGLAENKSNFVTSTTILIFCFIL-
{'generated_text': 'M\nRRAVGNADLGMEAARYEPSGAYQASEGDGAHGKPHSLPFVALERWQQLGPEERTLAEAVR\nAVLASGQYLLGEAVRRFETAVAAWLGVPFALGVASG-
```

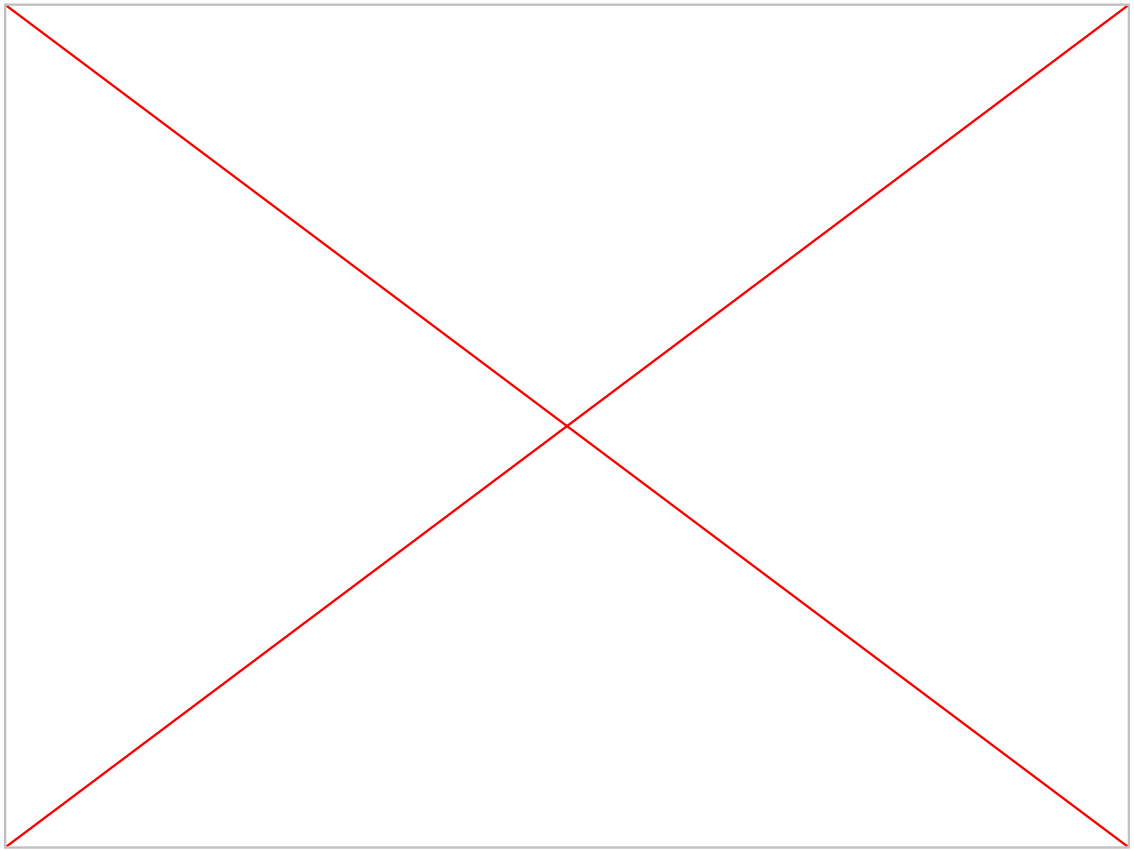
Evaluation of model

- Comparison result by: AlphaFold predictions, Rosetta Relax scores, molecular dynamics simulations.

	Natural dataset	ProtGPT2 dataset
IUPred3 (globular domains)	88.40%	87.59%
Ordered content	79.71%	82.59%
Alpha-helical content	45.19%	48.64%
Beta-sheet content	41.87%	39.70%
Coil content	12.93%	11.66%

($n = 10{,}000$ independent sequences/dataset).

Conclusion



Thank you for your attention

“We stop being observers of molecular life and start actively participating in the creation of new beneficial proteins.”

Resourses

- WANG, Lei; LI, Xudong; HAN, Zhang; WANG, Jinyi; JIANG, Di et al. A Comprehensive Review of Protein Language Models. Online. *ArXiv (Cornell University)*. 2025. Dostupné z: https://doi.org/10.48550/arxiv.2502.06881. [cit. 2025-12-08].
- FERRUZ, Noelia; SCHMIDT, Steffen a HÖCKER, Birte. ProtGPT2 is a deep unsupervised language model for protein design. Online. *Nature Communications*. 2022, vol. 13, no. 1, s. 4348-4348. ISSN 2041-1723. Dostupné z: https://doi.org/10.1038/s41467-022-32007-7. [cit. 2025-12-08].