On the *Self-supervised Learning* of protein engineering

Boyuan Wang

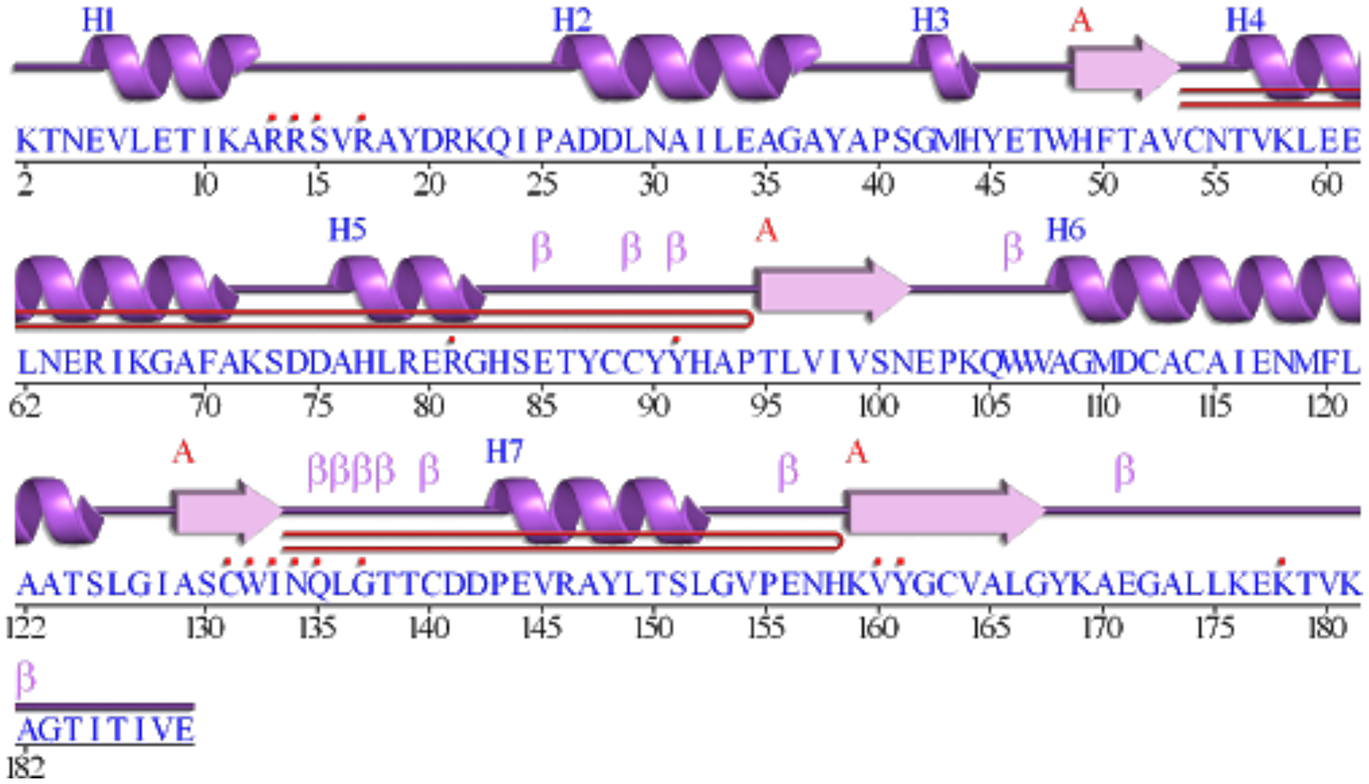
2020-04-23

What is protein engineering?

- **Protein engineering** is the process of developing useful or valuable proteins. It is a young discipline, with much research taking place into the understanding of protein folding and recognition for protein design principles. -*from Wikipedia*
- Common tasks in protein engineering:
  - Secondary structure prediction (1D)
  - Contact map prediction (2D)
  - Protein folding prediction (3D)

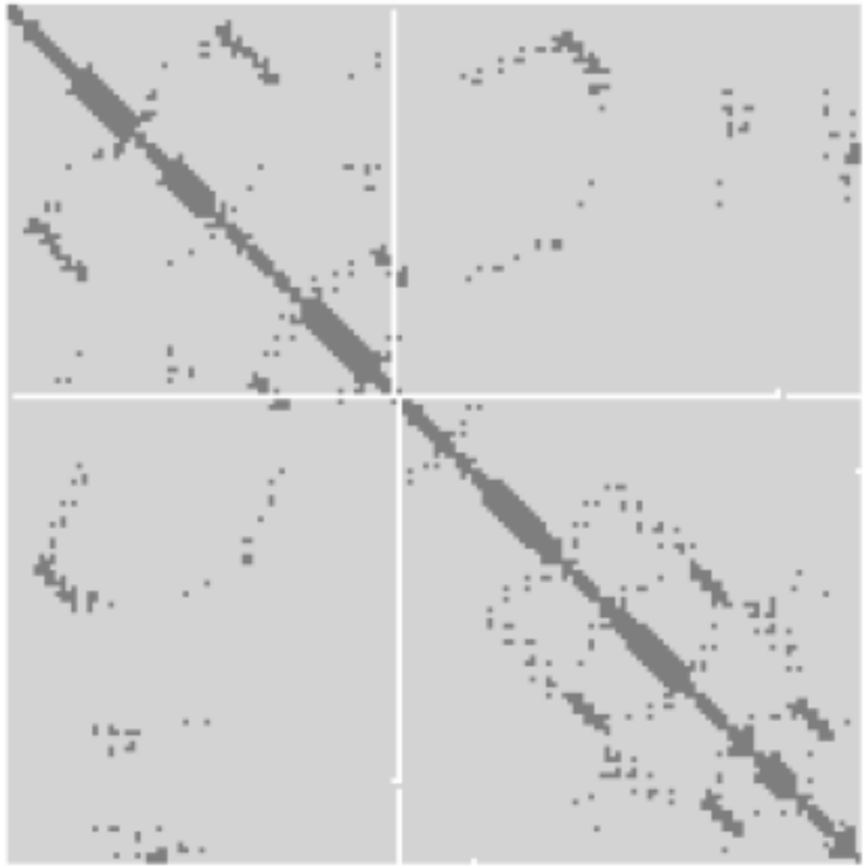
Secondary Structure Prediction

- Predict the position of alpha-helix (H) and beta-strand (E), coil region(C).



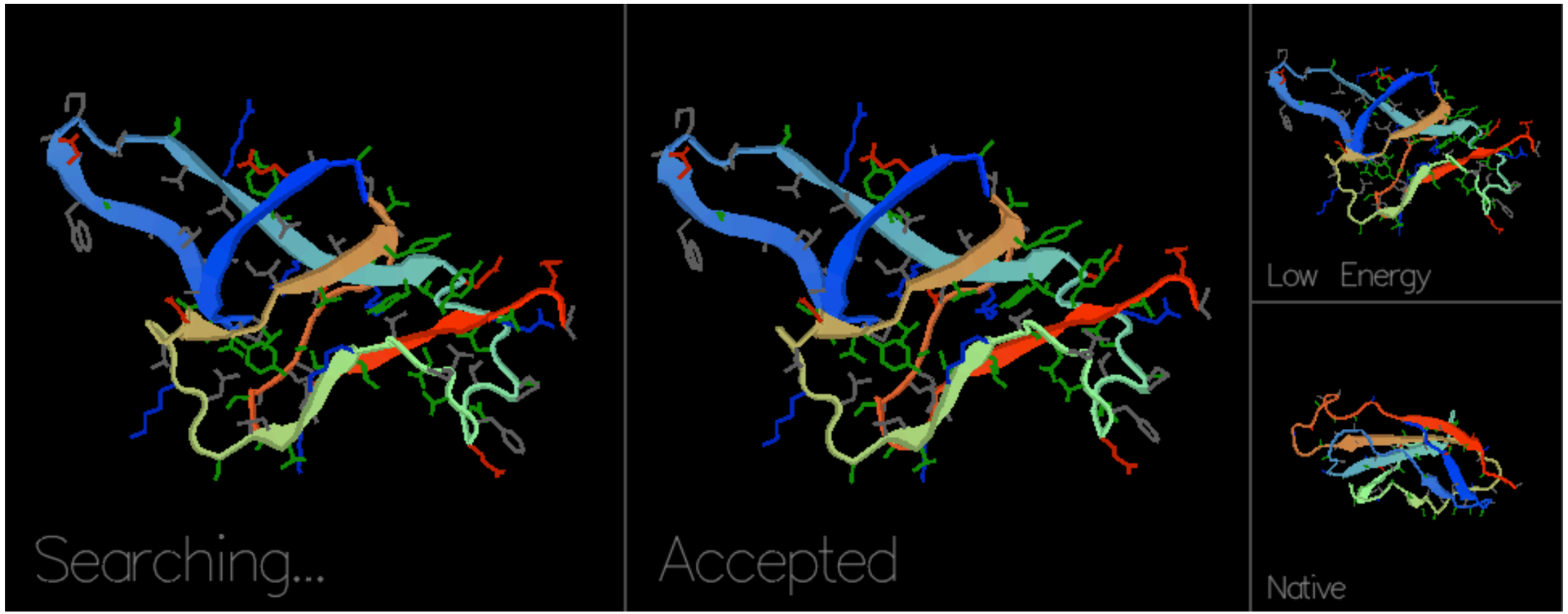
Contact map prediction

- Predict the contact information of amino acid residue



Protein folding prediction

- Predict the 3D geometric folding shape of proteins like Google Alpha-fold. (Hardest)



Why Do We Care Self-supervised Learning?

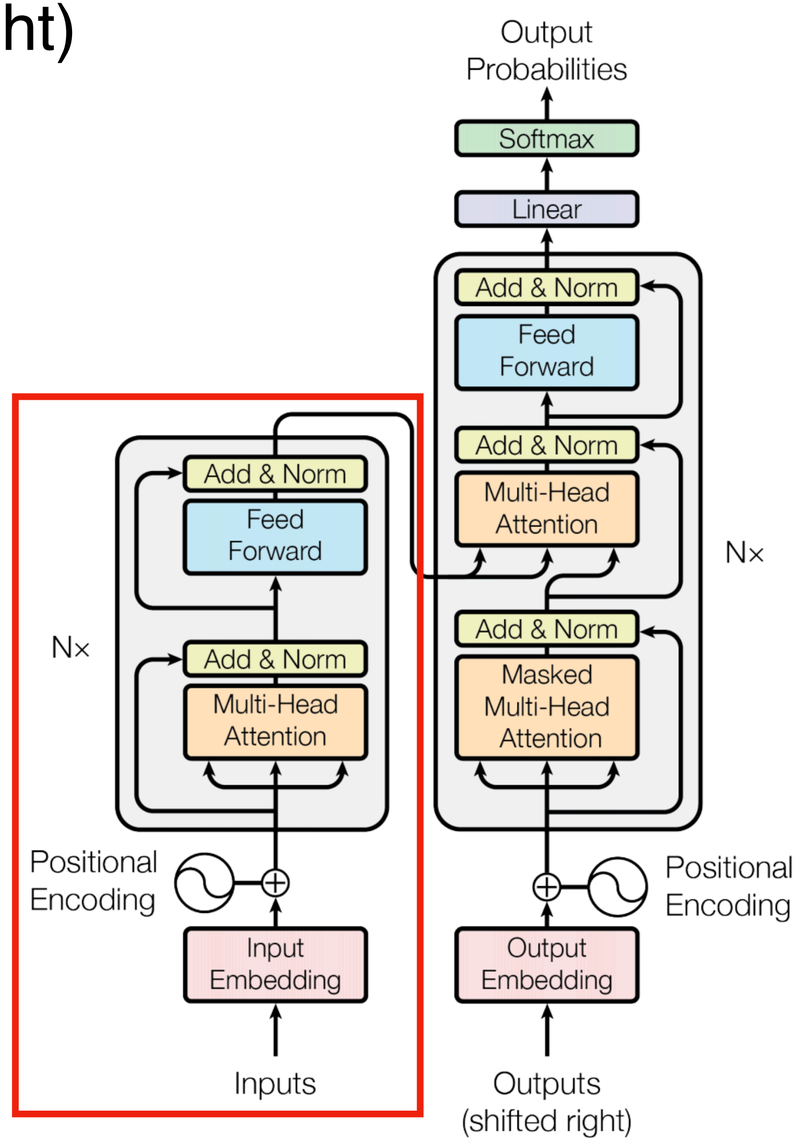
- Old methods involves too much human engineering work from selecting features to define functions for specific tasks.
- Recent use of deep supervised learning in protein engineering alleviates human laboring and brings exciting improvement in many tasks.
- However, data is scarce and obtaining supervised dataset is extremely costly in protein domain.
- Unlabelled protein data is abundant and contains the fundamental knowledge of proteins.
- Self-supervised learning is able to utilize the massive unlabelled data and extract knowledge from it.

Overview

- BERT: An Brief Introduction
  - Bidirectional Encoder Representations from Transformers, which is a pertained masked language model.
- Unified rational protein engineering with sequence-based deep representation learning (Nature Method 2019)
  - Rational protein engineering requires a holistic understanding of protein function. This paper proposed to use RNN based model to learn the holistic knowledge of protein sequences.
- Evaluating Protein Transfer Learning with TAPE (NeurIPS 2019)
  - This paper implements a more extensive framework by training three different self-supervised models. It also provided benchmark results on 5 standard tasks.
- ~~Generative models for graph-based protein design (NeurIPS 2019)~~
  - ~~This paper introduce a conditional generative model for protein sequences given 3D structures based on graph representations. (incomplete)~~

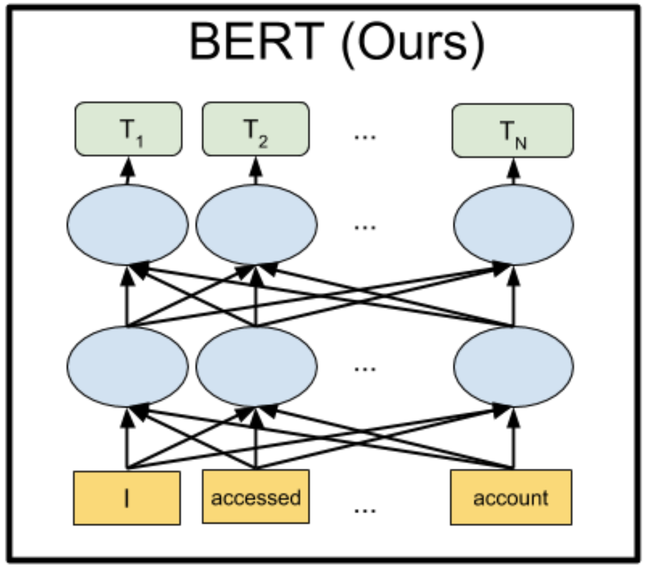
BERT - Architecture

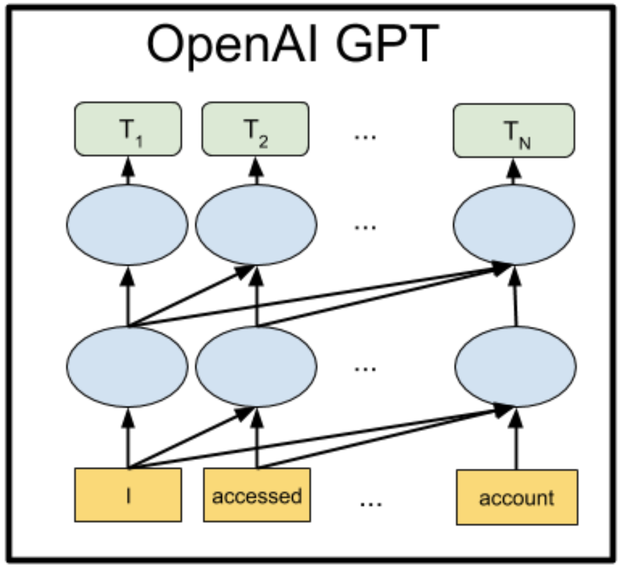
- A stack of Transformer Encoder. (red box in the right)
- Bidirectional representation.

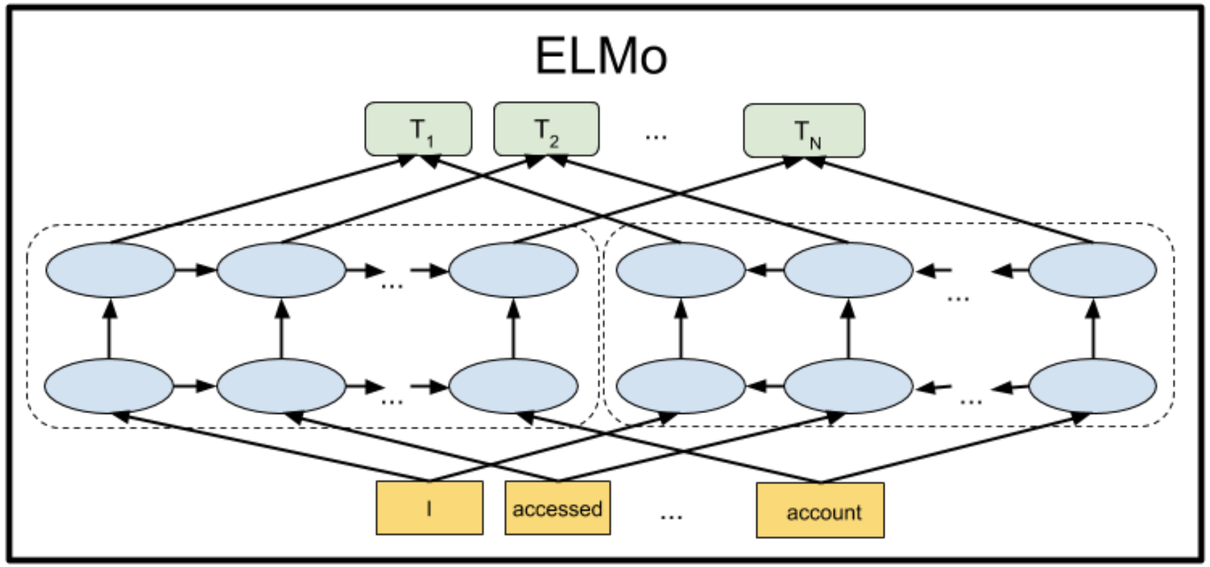


BERT - Architecture

- A stack of Transformer Encoder.
- Bidirectional representation.

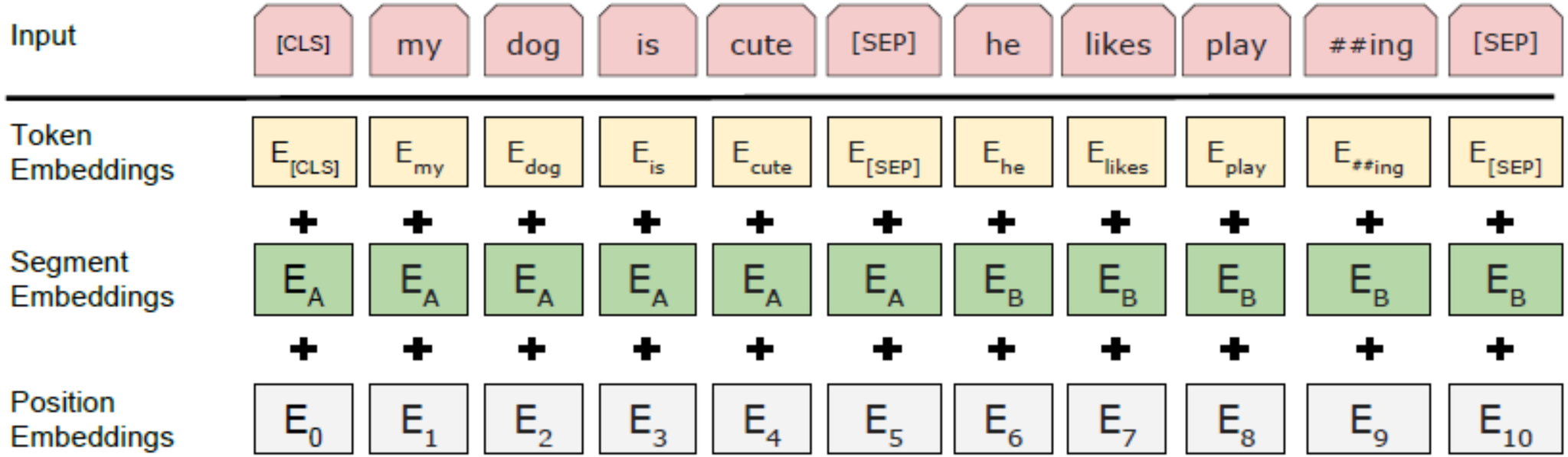






BERT - Input Features

- Token embedding + position embedding +. Segment embedding (sentence pairs)

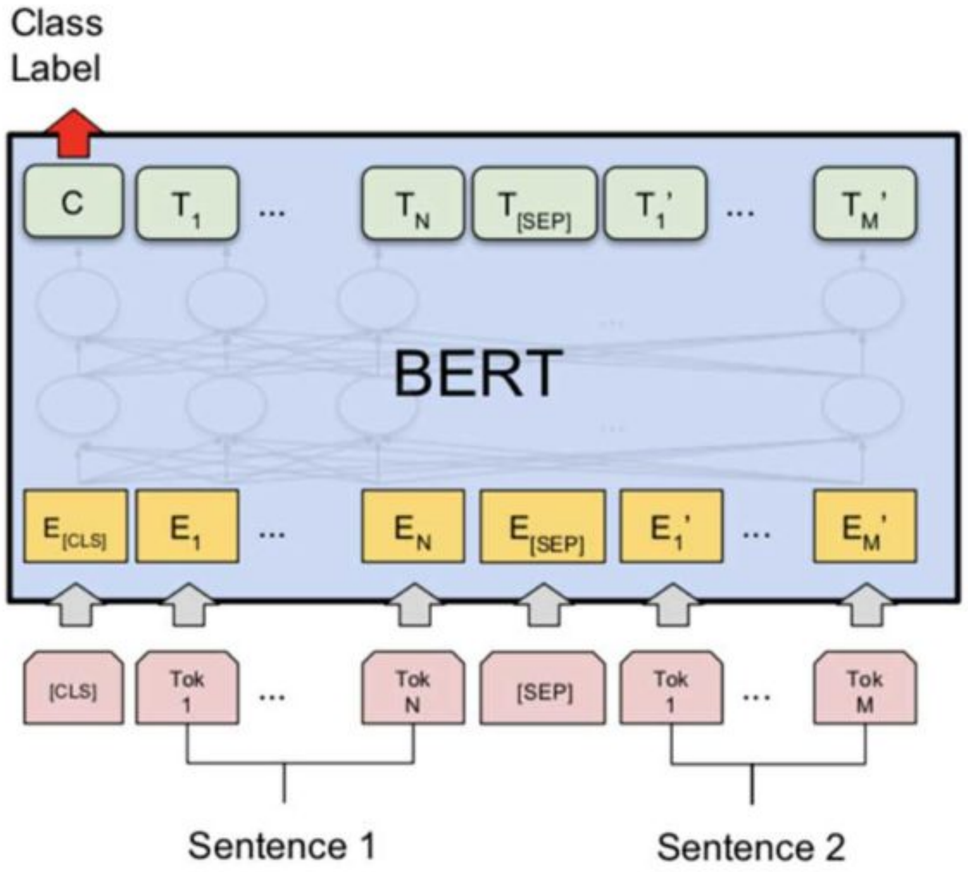


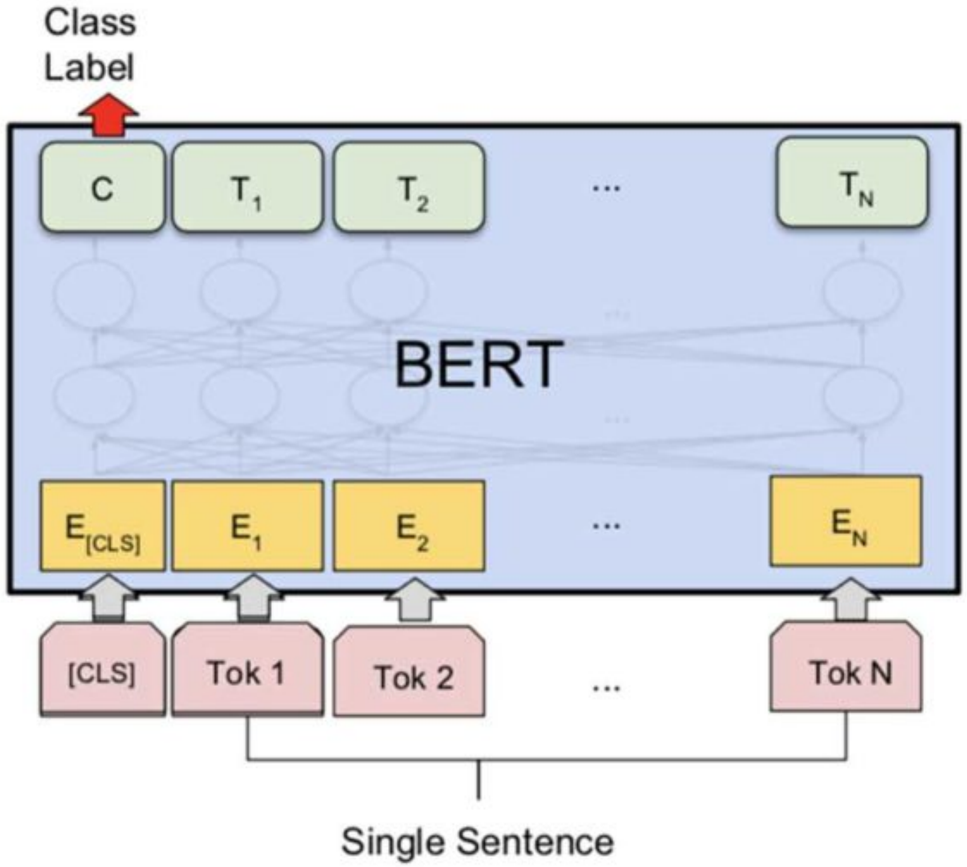
BERT - Pretrain Task

- Masked Language Model (MLM)
  - Mask 15% of tokens. Amount this 15%, 10% replaced, 10% unchanged.
  - 80%：my dog is hairy -> my dog is [mask]
  - 10%：my dog is hairy -> my dog is apple
  - 10%：my dog is hairy -> my dog is hairy
- Next Sentence Prediction (NSP)
  - Input sentence pairs (A, B), 50% of time B is the next sentence of A.
  - For question answering and natural language inference.

BERT - Fine-Tuning

- Fine-tuning on your specific tasks.
- [CLS] or token-level representation.



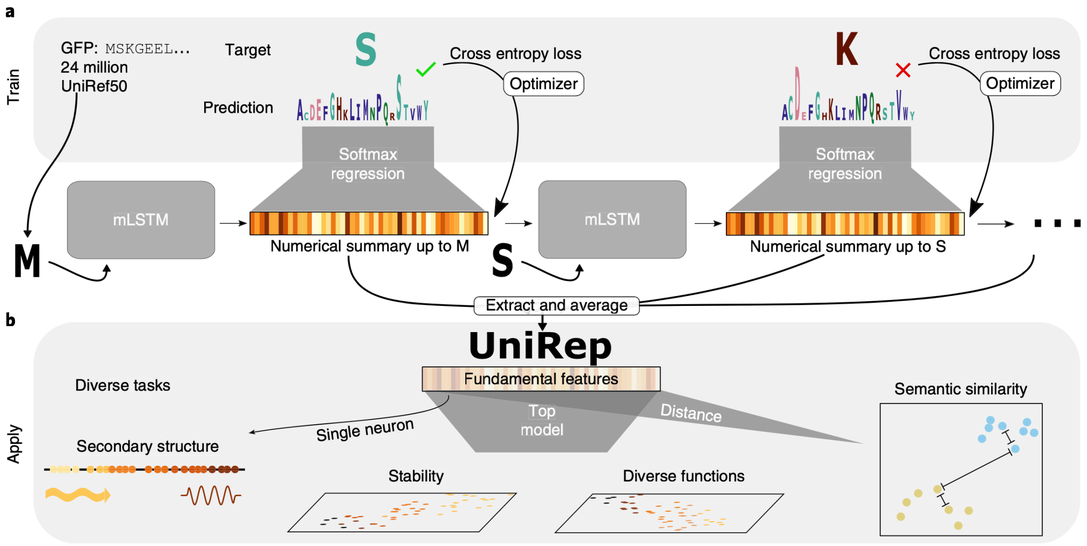


Paper 1 - Motivation

- Proteins are sequential data. We need to consider the spatial-temporal relationship.
- Likewise, natural language process (NLP) also deal with sequential data. We can adopt the algorithms from NLP domain to protein domain.
- Self-supervision brings significant improvement in many NLP tasks because it learns some fundamental knowledge of language. It should also be the case for proteins.

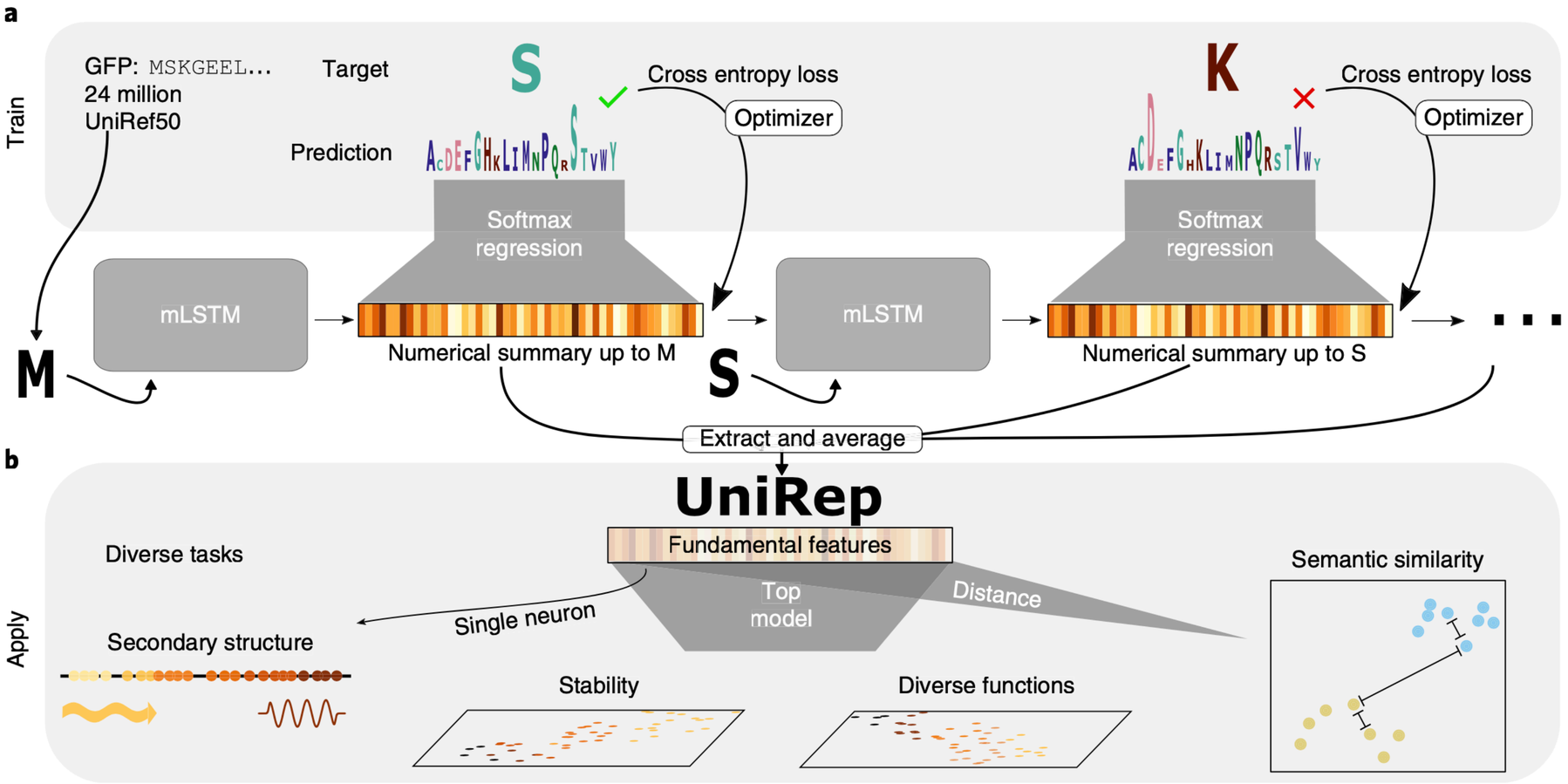
Paper 1 - Approach

- Self-supervision setup:
  - Architecture:
    - LSTM
    - Single-layer, 1900 hidden-size
  - Objective: Language model (next token prediction)
  - Data:
    - UniRef: ~24 millions sequence
    - Dictionary size: 20
  - Training Time:
    - ~770K steps, 1 epoch



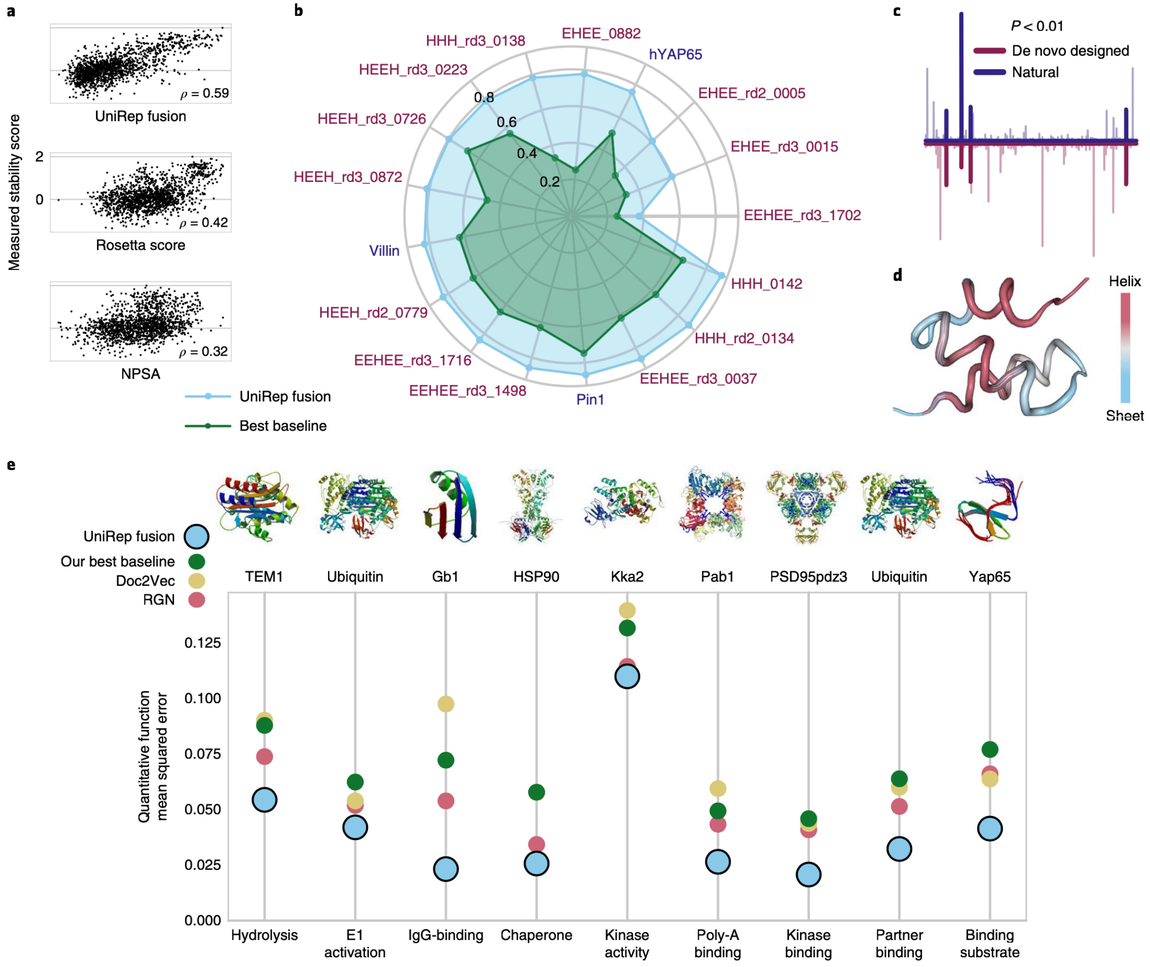
Paper 1 - Approach

- Training process:
  - Self-supervision: language modeling.
  - Downstream tasks: supervised learning.



Paper 1 - Experimental Results

- UniRep Feature:
  - averages all hidden states across time axis to make it more longterm dependent.
- Some results:



Paper 1 - Conclusion

- UniRep learns from raw data.
- It is unconstrained by a specific task, so features can be used in many tasks.
- It shows that protein informatics can potential go well directly from sequence to design.

Paper II - Motivation

- The first attempt for systematically evaluating semi-supervised learning on protein sequences.
- TAPE includes a set of five biologically relevant supervised tasks that evaluate the performance of learned protein embeddings across diverse aspects of protein understanding.
- A framework for multi-tasks benchmark.

Paper II - Tasks

- **Task 1**: Secondary Structure (SS) Prediction
  - Impact: understanding the function of a protein. Important for high level of structure prediction.
- **Task 2**: Contact Prediction
  - Impact: global information. Important for final 3D structure prediction.
- **Task 3**: Remote Homology Detection
  - Type: multilabel classification
  - Impact: detection of emerging antibiotic resistant genes and discovery of new enzymes.
- **Task 4**: Fluorescence Landscape Prediction
  - Type: regression
  - Impact: efficient exploration of the landscape.
- **Task 5**: Stability Landscape Prediction
  - Type: regression
  - Impact: important to ensure that drugs are delivered before they are degraded.

Paper II - Datasets

Table S1: Dataset sizes

Task	Train	Valid	Test
Language Modeling	32,207,059	N/A	2,147,130 (Random-split) / 44,314 (Heldout families)
Secondary Structure	8,678	2,170	513 (CB513) / 115 (TS115) / 21 (CASP12)
Contact Prediction	25,299	224	40 (CASP12)
Remote Homology	12,312	736	718 (Fold) / 1,254 (Superfamily) / 1,272 (Family)
Fluorescence	21,446	5,362	27,217
Stability	53,679	2,447	12,839

Paper II - Approach

- Self-supervised Learning Setup:
  - LSTM (RNN)
    - forward 3-layer LSTM+ backward 3-layer LSTM, 1024 hidden size.
    - loss: language modeling + task fine-tune.
  - Bert (SAN)
    - 12-layer, 512 hidden size, 8 attention head.
    - loss: Masked language modeling + fine-tune.
  - ResNet (CNN)
    - 35*(2 conv-layer with 256 filter), kernel size 9, dilation rate 2.
    - loss: language model + fine-tune.

Paper II - Experiment Results

Table 1: Language modeling metrics

	Random Families			Heldout Families		
	Accuracy	Perplexity	ECE	Accuracy	Perplexity	ECE
Transformer	**0.45**	**8.89**	**6.01**	**0.30**	**13.04**	**10.04**
LSTM	0.40	**8.89**	6.94	0.16	14.72	15.21
ResNet	0.41	10.16	6.86	0.29	13.55	10.32
Supervised LSTM [11]	0.28	11.62	10.17	0.14	15.28	16.02
UniRep mLSTM [12]	0.32	11.29	9.08	0.12	16.36	16.92
Random	0.04	25	25	0.04	25	25

Paper II - Experiment Results

Table 2: Results on downstream supervised tasks

Method		Structure		Evolutionary	Engineering	
		SS	Contact	Homology	Fluorescence	Stability
No Pretrain	Transformer	0.70	0.32	0.09	0.22	-0.06
	LSTM	0.71	0.19	0.12	0.21	0.28
	ResNet	0.70	0.20	0.10	-0.28	0.61
Pretrain	Transformer	0.73	0.36	0.21	**0.68**	**0.73**
	LSTM	0.75	0.39	**0.26**	0.67	0.69
	ResNet	0.75	0.29	0.17	0.21	**0.73**
Supervised [11]	LSTM	0.73	0.40	0.17	0.33	0.64
UniRep [12]	mLSTM	0.73	0.34	0.23	0.67	**0.73**
Baseline	One-hot	0.69	0.29	0.09	0.14	0.19
	Alignment	**0.80**	**0.64**	0.09	N/A	N/A

Paper II - Conclusion

- The improve over labelled data shows promising future for self-supervision in protein prediction.
- No single self-supervised model performs best across all protein tasks. Needs the extensive benchmark to evaluate the models.
- Structure prediction still is inferior to the alignment method. In need for better self-supervision design and studying the relationship between alignment and learned-based representation.