Project Report: ESM1b-e2e

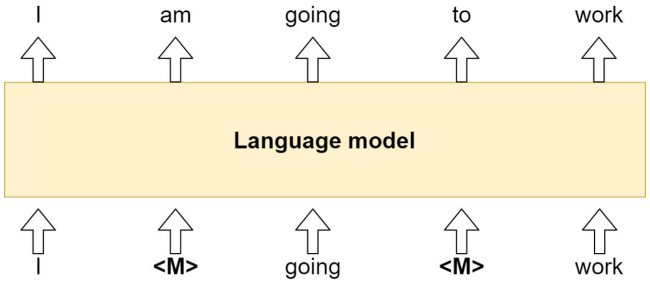
Sahasra Ranjan

Predicting Km and the Unirep Model

- **Masked Language Modeling:** Technique used in NLP for learning the language and representations for the given sequences

- **UniRep model**: Previous model used for Km prediction. It is based on LSTM which is slow to train

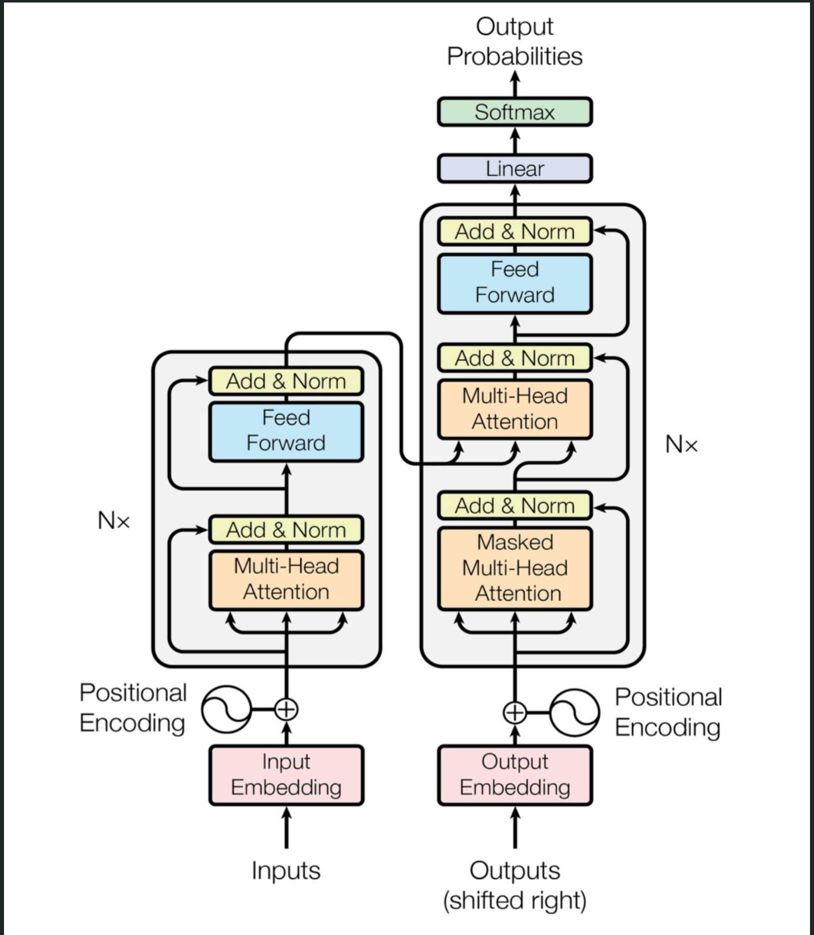
- We are now introducing the new model based on Transformer networks (current state of the art for NLP tasks)



Transformer Network

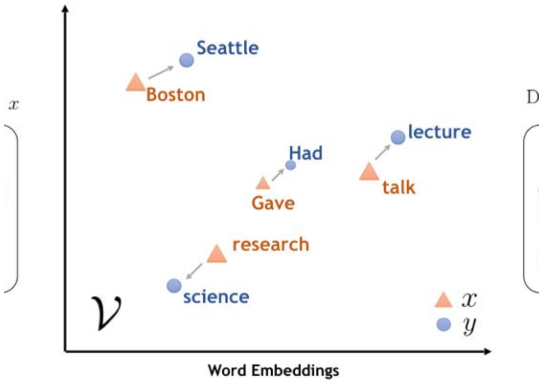
Attention is All You Need *(2017)*

- *Vaswani et. al.*



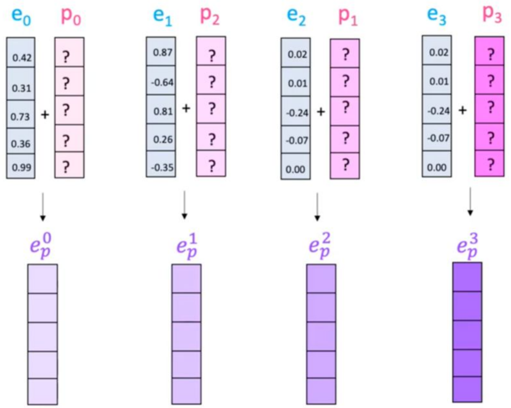
Token Embeddings

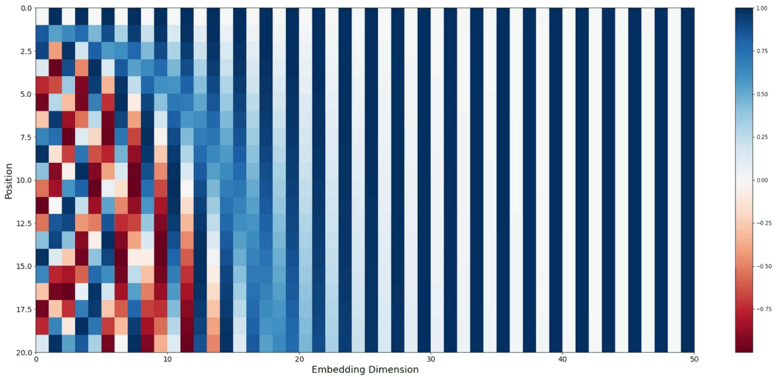
- Why do we need embeddings in the first place? Machines don't understand english languages, but matrices. So we want to have a **matrix representation/mapping** for in input language (protein sequences for our case).



- Transformer network takes all of these embedding at once for the input, so **positional embedding** was introduced to store the order of these embeddings in the original input.

Positional Embeddings





- We want to store information about the position, so why not just add something (and later subtract) to the original embedding. But what?

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{\text{model}}})$$
$$PE_{(pos,2i+1)} = cos(pos/10000^{2i/d_{\text{model}}})$$

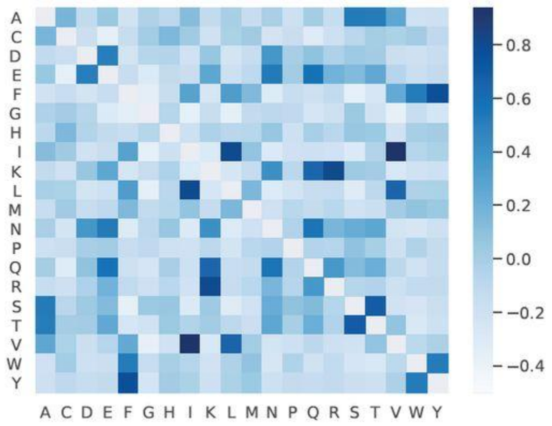
where, *pos* is the index in the input and *i* is the depth (d-dimensional)

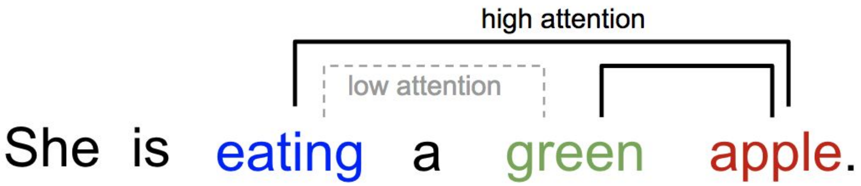
Self-Attention

Say the following sentence is an input sentence we want to translate:

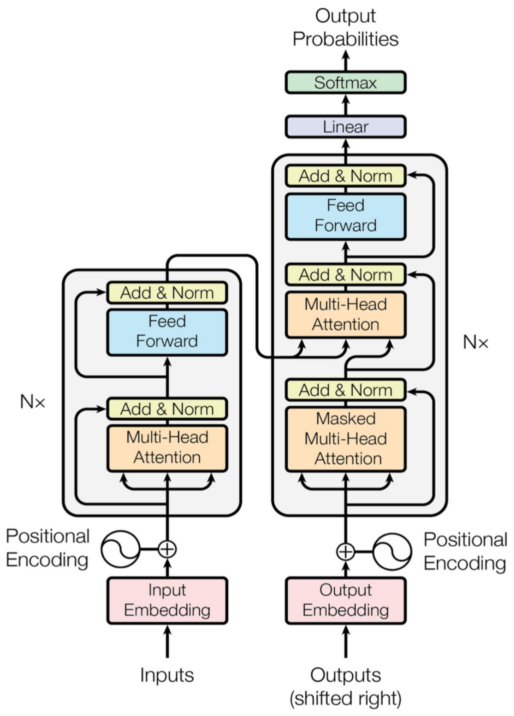
"She is eating a green apple"

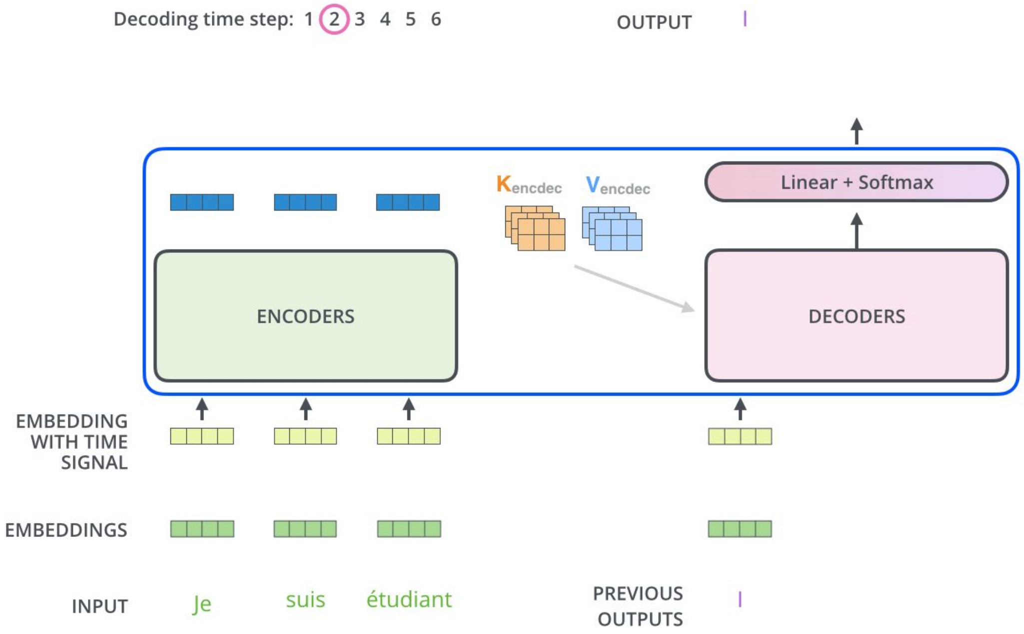
What does "eating" in the sentence refer to? Self-attention allows it to associate "eating" with "apple".



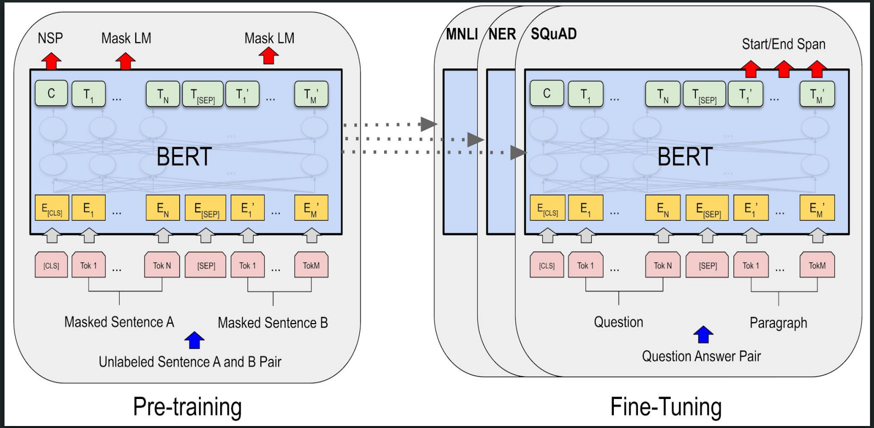


Encoder and Decoder





BERT

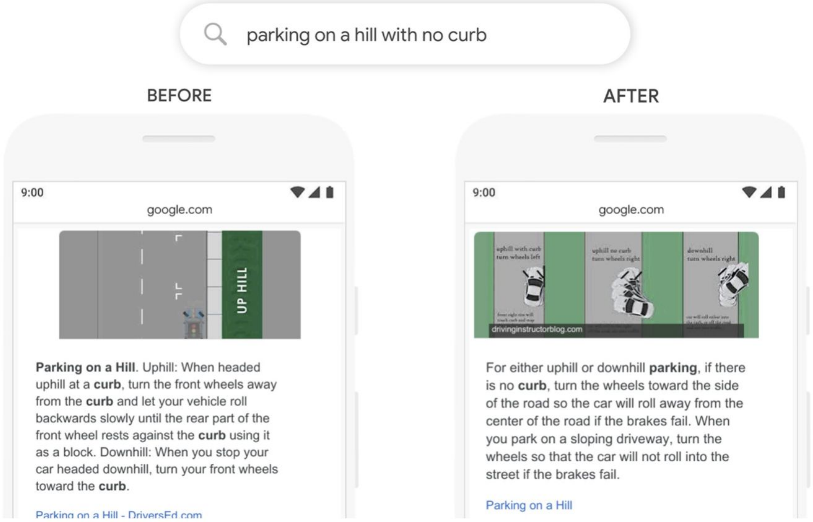


BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (*2018*)

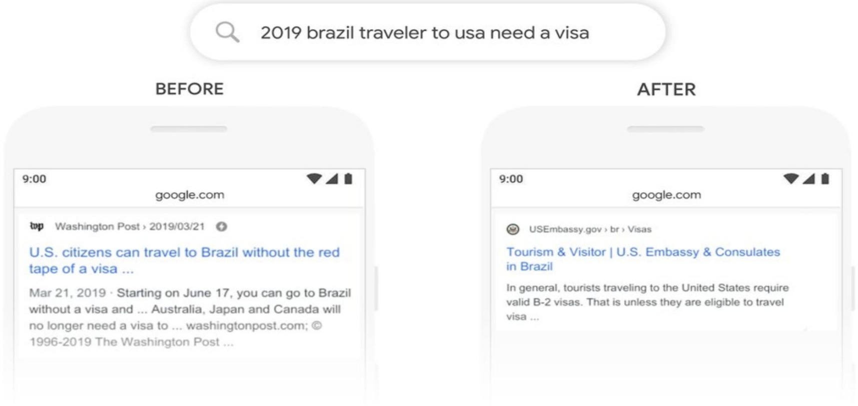
- Devlin et al.

Introduction

BERT model was first introduced by Google AI team, which helped them improve the Google search results for complex queries.



In the past, a query like this would confuse our systems--we placed too much importance on the word “curb” and ignored the word “no”, not understanding how critical that word was to appropriately responding to this query. So we’d return results for parking on a hill with a curb!

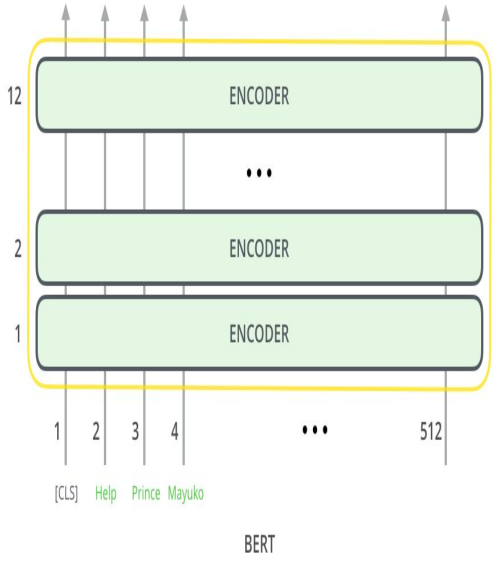


BERT vs Transformers

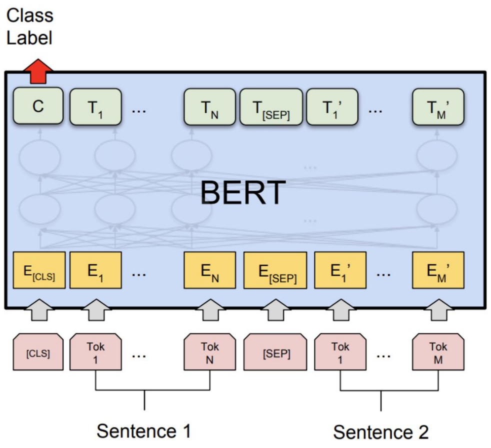
- BERT has multiple **encoder stacked above one another** whereas Transformer uses two separate stacks, Encoders and Decoders which are connected to each other.

- BERT model are pre-trained and the fine-tuning for specific task gives much better result in lesser training time as compared to Transformers.

- It’s not always to case that the pretrained BERT model is available (open-sourced) for the specific task that is required.



Classification tasks using the BERT model



- The [CLS] token!!
- The first input token in the BERT model is a <start> token i.e. [CLS].
- This can be used for classification tasks as the output corresponding to [CLS] is a *d*-dimensional vector which stores information about the complete input sequence.

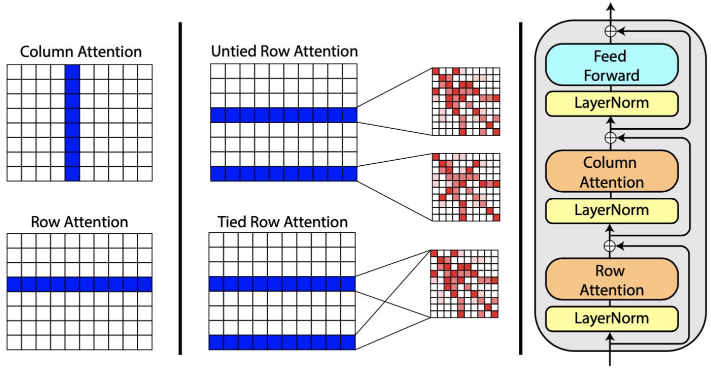
ESM1b - Protein BERT

- *Rives et al.*

- **ESM1b-BERT** model for proteins was proposed by the FAIR (Facebook AI Research) team which can be used general-purpose protein language modeling

- The FAIR team provided pre-trained weights for the Protein BERT model which we used after fine-tuning the weights for our task of Km prediction

The MSA transformer: Introduced by the Facebook AI Research team, in the first quarter of 2021



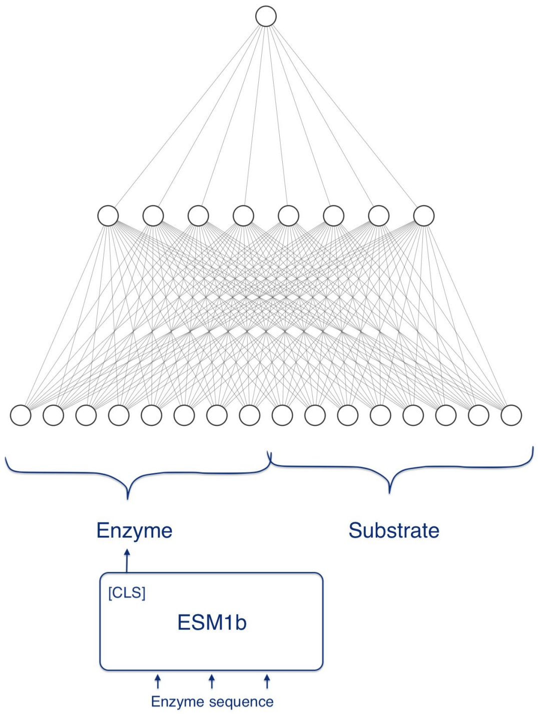
Enzyme-Substrate Binding

- Given a pair of enzyme and substrate, we want to predict if they will bind during a reaction or not

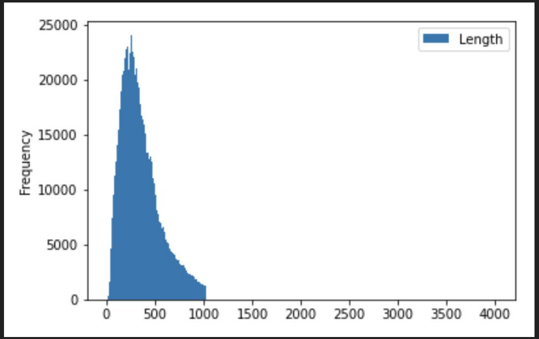
- For this task, we used the ESM1b model and added a **fully-connected layer** with input as enzyme vector (extracted from ESM1b) and substrate vector.

- We trained the fully-connected layer to predict if the enzymes and substrate bind together or not.

- This model not only learns the representation but also the positional information which improved the accuracy of the predictions



Dataset



- We used the **UniProt-50 dataset** for the language modeling task

- Uniprot-50 dataset has close to 33 million protein sequences (20GBs). Out of these we extracted those which are enzymes (around **3 million**, 2GBs)

- Since, ESM1b model can take sequences of length at max 1024 so we split the sequences with length longer than 1023 tokens

Data processing scripts:

- We used AWK scripts for processing and extracting data-points from the original UniProt dataset
- These scripts might be handy for other projects related to fasta sequences

Scripts:

- **extract.sh**: Split fasta file in the three sub-parts (train, test, val) for given splits
- **get_len.sh:** Read the fasta file and return the length of each sequence in same order
- **head_seq.sh:** Extract a-th to b-th sequences from the original file given ‘a’ and ‘b’
- **shorten.sh:** Extract sequences with length less than the given max length and crop large sequences to fit the max length parameter
- **shuffle_select.sh:** Extract given number of sequences (at a step of total/num)

Repo link to the scripts

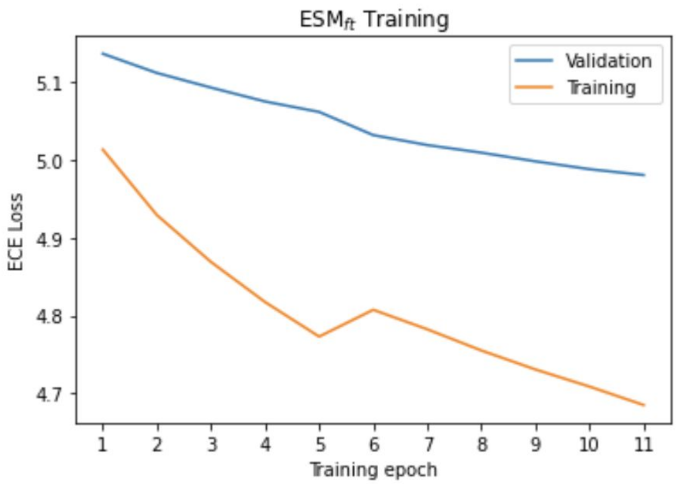
Training

- Devices used:
    - Nvidia A100
        - 40 GB RAM
    - Nvidia RTX-6000
        - 24 GB RAM

- We trained our final model for 10 epochs on 8 GPUs together by distributing the data across the GPUs (Distributed Data Parallel).

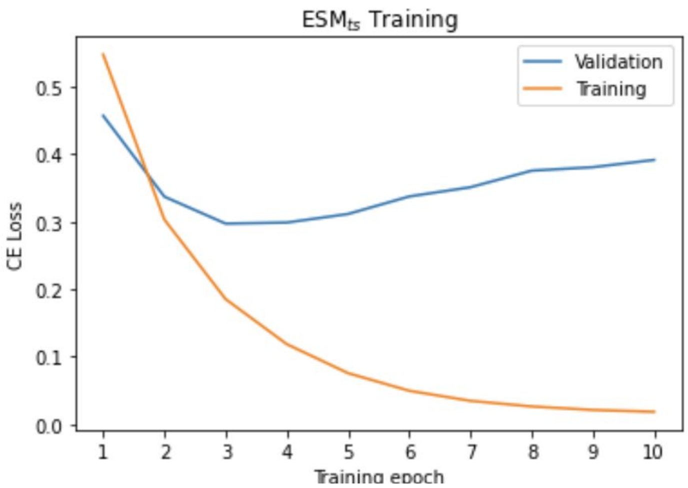
- Training our final model took around 100 hrs to complete on 8 A100 GPUs

Results



ESM1b$_{ft}$

Started with a val ECE loss of 5.2138



ESM1b$_{ts}$ (e2e)

References:

- Attention Is All You Need - *Vaswani et al.*
- Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences - *Rives et al.*
- BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding - *Devlin et al.*
- MSA Transformer - *Rao et al.*
- hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment - *Remmert et al.*
- The Illustrated Transformer - *Jay Alammar*

Thank You!!