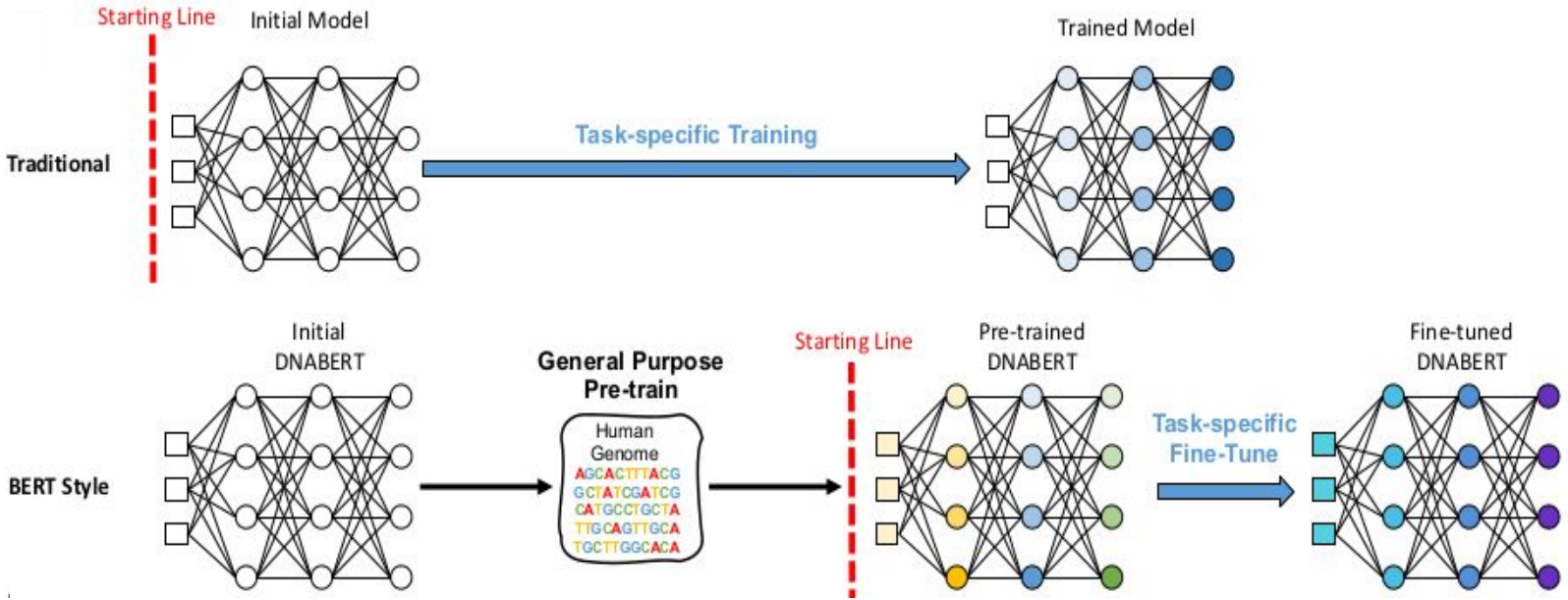
Fast Pretraining of Gene Sequences to Enable Few-Shot Learning

Soumyadeep Roy, Niloy Ganguly (IIT Kharagpur, India)
Jonas Wallat, Sowmya S Sundaram, Wolfgang Nejdl
(L3S Research Center, Germany)

26th European Conference on Artificial Intelligence (ECAI 2023)

Genomic Pretraining

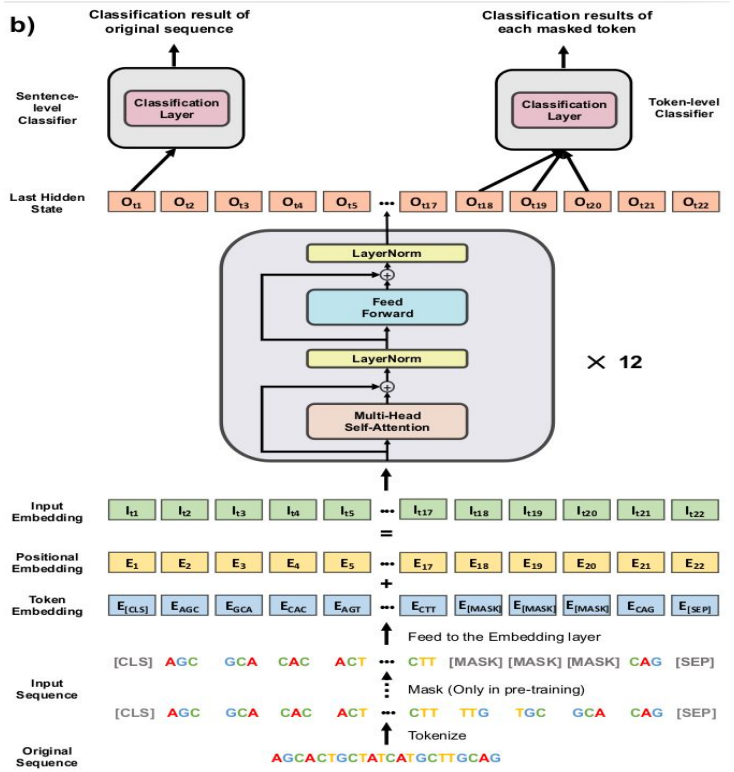


Ji et al. (2021) DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, Bioinformatics, pp. 1-9

Base model - DNABert

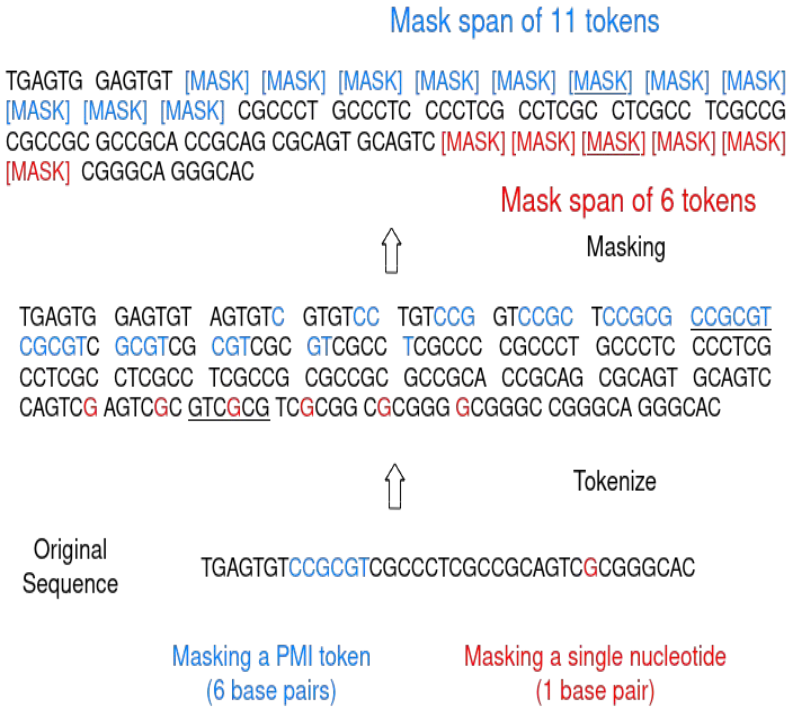
- DNABert - Combines BERT with DNA sequences of 6-mers (ATTCGC)
  - Model vocabulary size for 6-mer model: 4^6 plus special tokens

- Gene regulatory code (non-coding) is complex, shows signs of polysemy, distant semantic relationship between sequence codes
  - Cis-regulatory elements acts similar to language

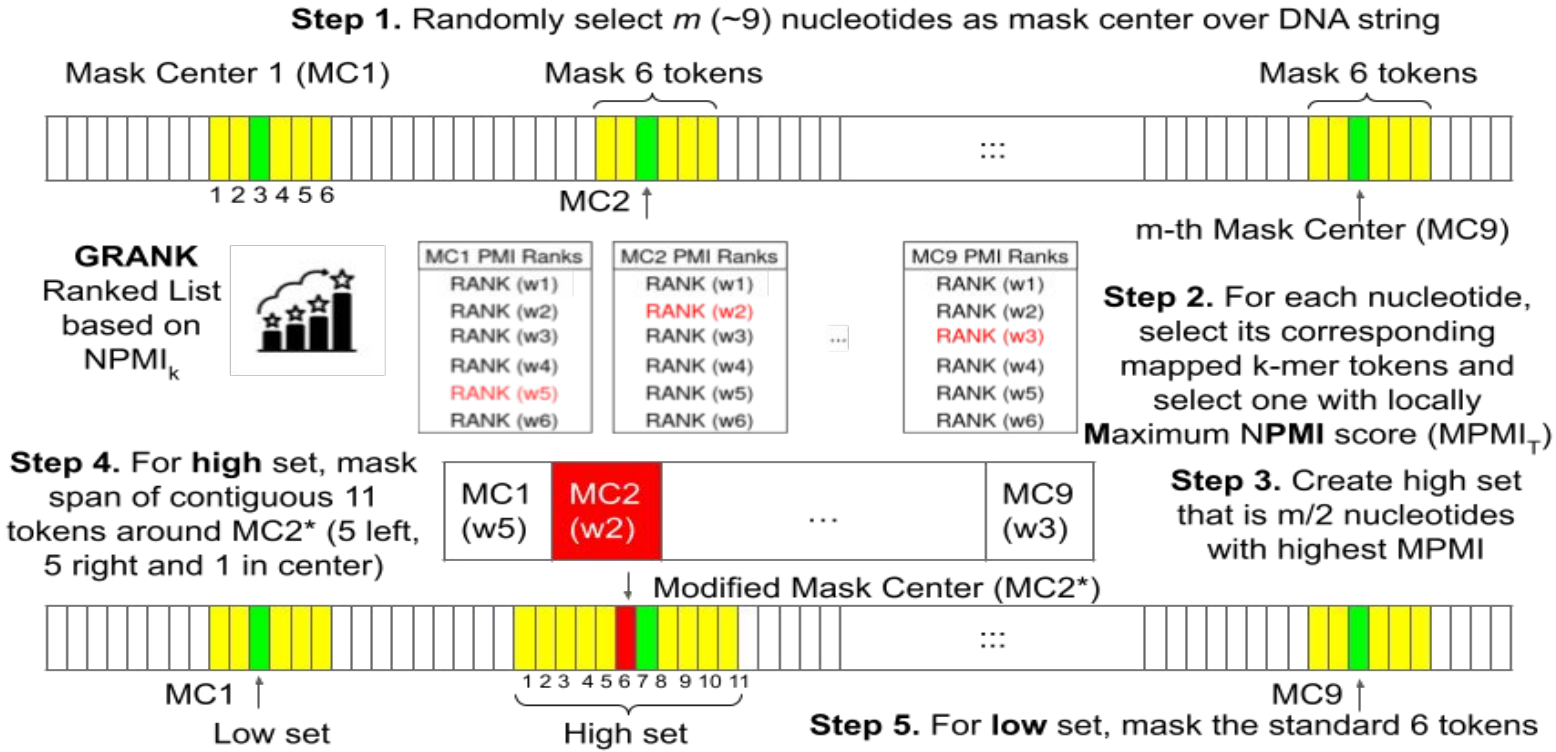


GeneMask (PMI-based) Masking

- Random masking allows abusing local features (not learning the overall context)
  - the United [MASK] : the United States
  - by [MASK] way : by the way
  - Training steps are wasted for “easy” predictions
- Words in NLP = _____ in gene sequences
  - Difficult to identify semantic-preserved tokens
- Idea: Jointly mask multiple tokens if they exhibit high collocation



PMI Masking - mask longer correlated spans together



Downstream Tasks

- Promoter Region Prediction - binary classification
  - Prom-core: -35 bp to +34 bp around TSS
  - Prom-300: -249 bp to +50 bp around TSS

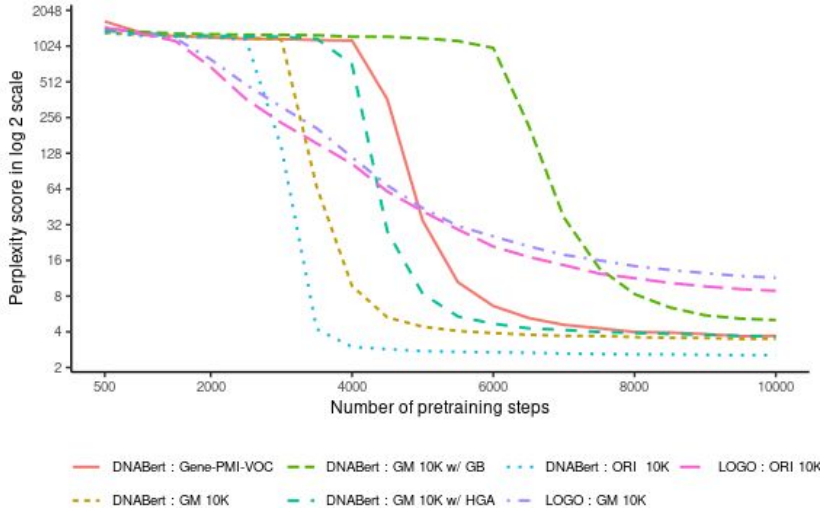
- Enhancer prediction - 500 bp
  - An enhancer is a sequence of DNA that can bound specific proteins and therefore increase a change of transcription of a particular gene. Unlike promoters, enhancers do not need to be in a close proximity to TSS (might be several Mb away)

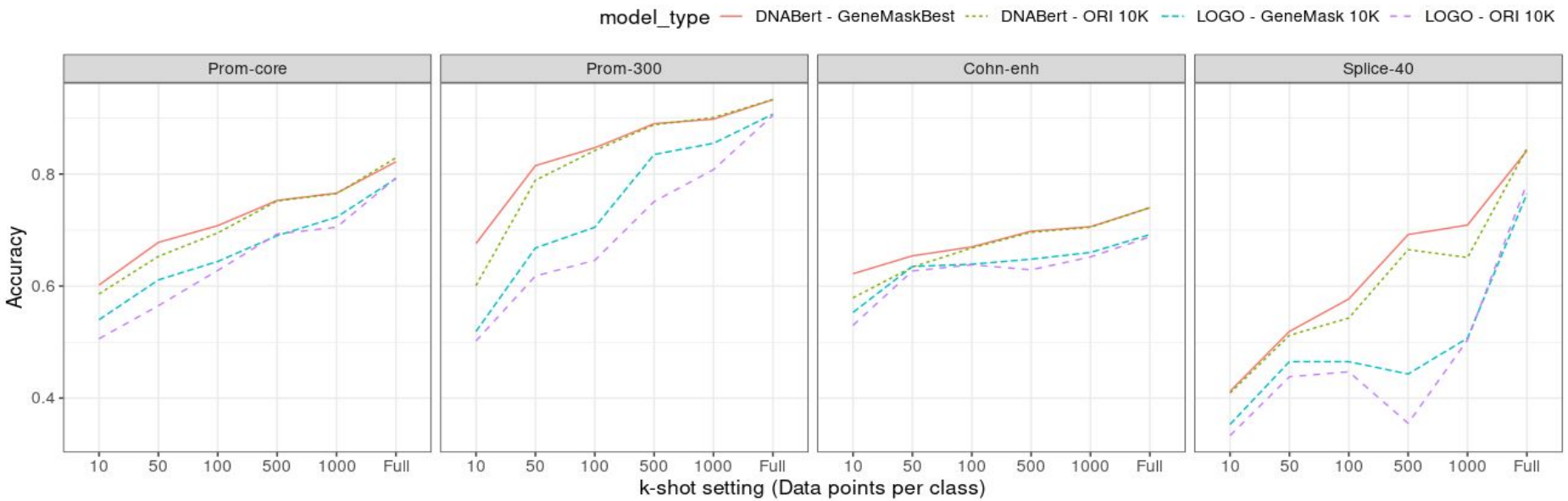
- Splice Donor and Acceptor Site Prediction - predict whether donor, acceptor or non-splice site (3-way classification) - 40 bp
  - Extract 40 bp long sequence around the donor and acceptor sites of exons as positive sequences

Experimental Setup

- Five few-shot settings - 10, 50, 100, 500 and 1000-shot
- Report mean accuracy and AUC from running 10 times with different seeds and fine-tuning data

- All baseline models are run for 10000 steps, which takes DNABert 2.5 days and LOGO 20 hours on 4 GTX1080Ti 11GB GPUs
    - ○Perplexity score converged to a low score and stable over last 3000 steps
    - ○Original DNABert paper trained for 120K steps





- Performance improvement is higher in LOGO (light-weight) than DNABert due to GeneMaskBest
- Performance improvement is highest in 10-shot, followed by 50-shot. Improvement diminishes at higher data settings
- GeneMask improves decently over DNABert models trained on 120K steps (ORI-120K) in most settings, except for Prom-core (10, 50 and 100-shot)

Conclusion

GeneMask ensures substantial speedup of 10x and performance improvement over random masking strategy of SoTA models (DNABert and LOGO) in few-shot settings

Incorporating domain knowledge while pretraining needs to be designed based on the (target) downstream task

Preprint available at: https://arxiv.org/abs/2307.15933

Code and data files available at: https://github.com/roysoumya/GeneMask