Diagnosing Cancer with Genomic Context

GENOMIC LANGUAGE MODEL | NOV.28.2024 | NAYOUNG KU | PROFESSOR TAEJIN AHN

genetic code and human language

What is NLP?

What is NLP?

“He works at Google.”

a company name

“He googles at work”

a verb meaning to search online

Protein Prediction

AlphaFold & RoseTTAFold
Computational Protein Design
Protein Structure Prediction

LLM

Meerkat-7B
The model learned 7B words is
a current SOTA in sLLM passing USMLE

Protein Prediction

AlphaFold & RoseTTAFold
Computational Protein Design
Protein Structure Prediction

LLM

Meerkat-7B
The model learned 7B words is
a current SOTA in sLLM passing USMLE

“Nucleotide Sequence as a Language”

“Diagnosable Pattern”

Agenda

PURPOSE

OBJECTIVES

METHODS

DISCUSSION

CONCLUSION

PURPOSE

To Explore the Application of NLP in Genomics for Cancer Diagnosis

#NLP #Context-Based Analysis #Cancer Diagnosis #Qualitative #Quantitative

OBJECTIVES

Text Preprocessing	Process RNA-seq data into gene-level amino acid sequences.
Embedding	Use ProtBERT to generate contextual embeddings of protein sequences.
Modeling	Develop deep learning models to classify cancerous vs. normal samples.
Validation	Validate the approach with ovarian cancer datasets and measure performance.

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

Summary for methods

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

Tokenization

RNA-seq data in FASTQ

TTAATGGTTAAATCCCT
ATACTAAGCA ...
(Quatlity Score)

Sampling:
- OVCAR: 100 / 144
- AC: 100 / 404

Extract the reads

Gene Level Sequence

TTA**ATGGTTAAATCCCT**
ATACTAAGCA

Token = Gene

Split the reads into gene-level sequence by identifying start and stop codons.

Translation

ATG/GTT/AAA/TCC/
CTA/TAC/TAA
MVKSLY

Split the reads into gene-level sequence by identifying start and stop codons.

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

ProtBERT: Pre-trained Language model

ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost

Abstract—Computational biology and bioinformatics provide vast data gold-mines from protein sequences, ideal for Language Models (LMs) taken from Natural Language Processing (NLP). These LMs reach for new prediction frontiers at low inference costs. Here, we trained two auto-regressive models (Transformer-XL, XLNet) and four auto-encoder models (BERT, Albert, Electra, T5) on data from UniRef and BFD containing up to 393 billion amino acids. The protein LMs (pLMs) were trained on the Summit supercomputer using 5616 GPUs and TPU Pod up-to 1024 cores. Dimensionality reduction revealed that the raw pLM-*embeddings* from unlabeled data captured some biophysical features of protein sequences. We validated the advantage of using the *embeddings* as exclusive input for several subsequent tasks: (1) a per-residue (per-token) prediction of protein secondary structure (3-state accuracy Q3=81%-87%); (2) per-protein (pooling) predictions of protein sub-cellular location (ten-state accuracy: Q10=81%) and membrane versus water-soluble (2-state accuracy Q2=91%). For secondary structure, the most informative embeddings (ProtT5) for the first time outperformed the state-of-the-art without multiple sequence alignments (MSAs) or evolutionary information thereby bypassing expensive database searches. Taken together, the results implied that pLMs learned some of the *grammar* of the *language of life*. All our models are available through https://github.com/agemagician/ProtTrans.

Index Terms—Computational biology, high performance computing, machine learning, language modeling, deep learning

https://ieeexplore.ieee.org/document/9477085
https://huggingface.co/Rostlab/prot_bert

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

ProtBERT: Pre-trained Language model

Protein-Specific Language Model

- Specific for protein sequences
- **The embeddings** generated by ProtBERT have shown **superior performance** in tasks such as predicting protein secondary structure and sub-cellular location.

Trained by 2 Datasets

UniRef100 & Big Fantastic Database (BFD)
> 393 billion amino acids

Self-Supervised Learning

Masked language modeling approach like **BERT**

- Understand the **grammar** of protein sequence
- Enhance its predicative capability of various biological tasks

METHODS

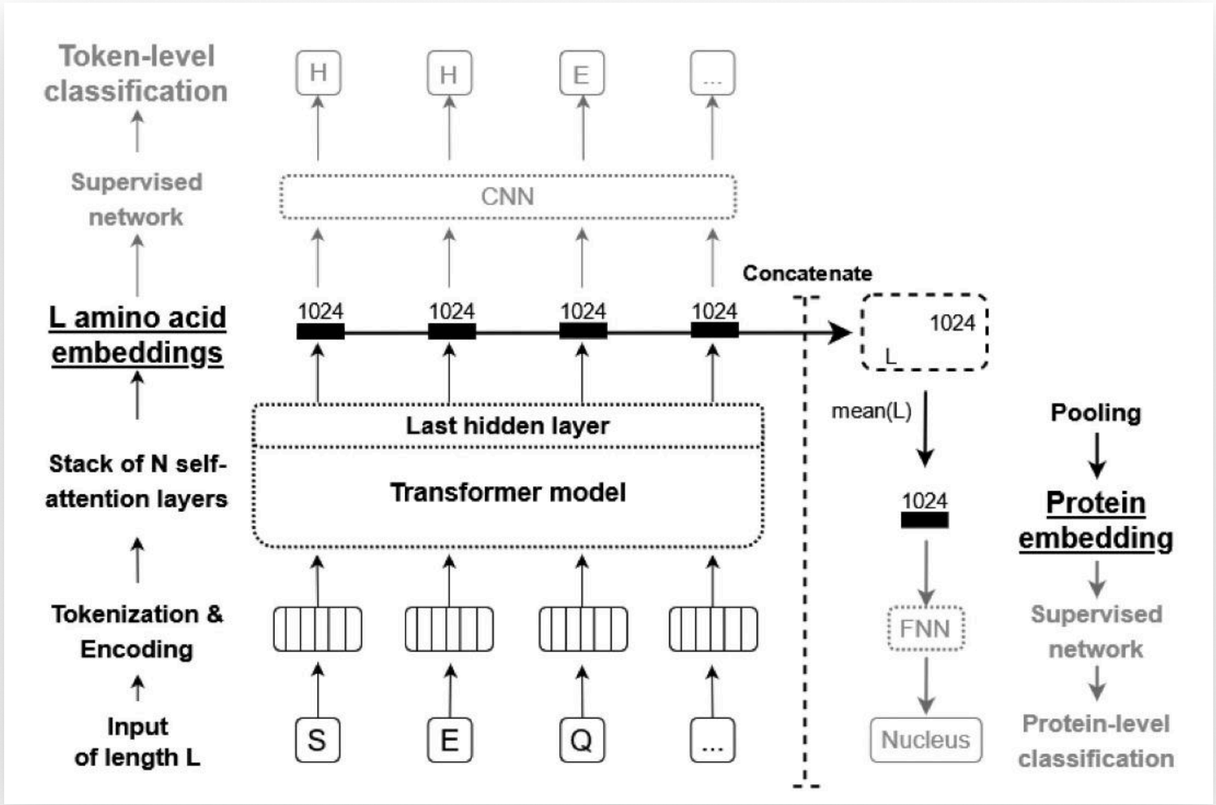
Text Preprocessing

Embedding

Modeling

Training & Validation

ProtBERT



https://ieeexplore.ieee.org/document/9477085
https://huggingface.co/Rostlab/prot_bert

METHODS

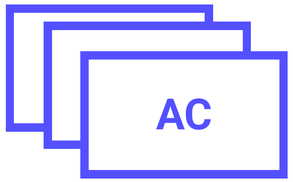
Text Preprocessing

Embedding

Modeling

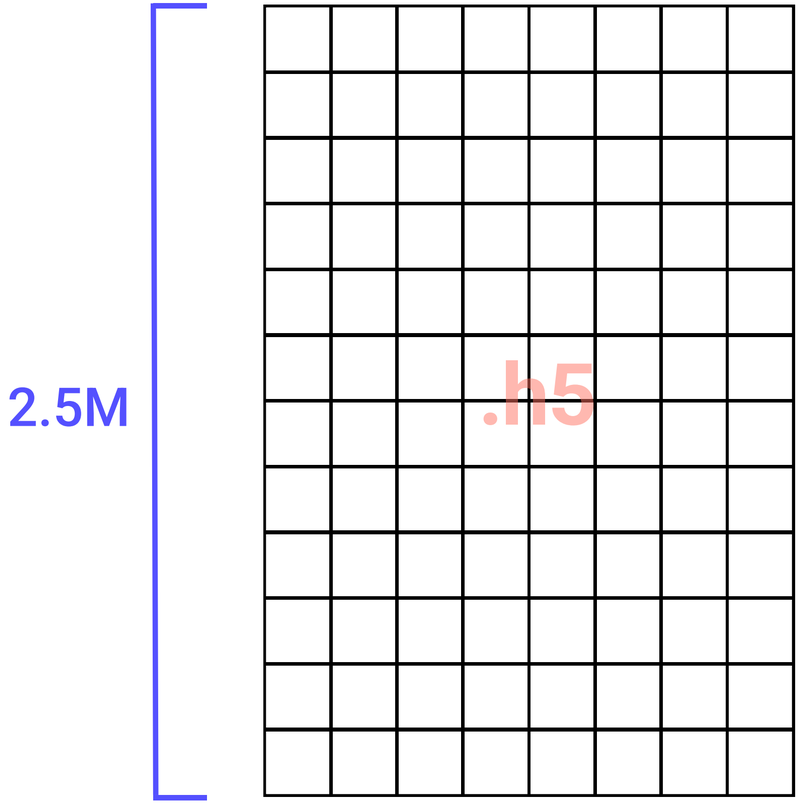
Training & Validation

n = 10

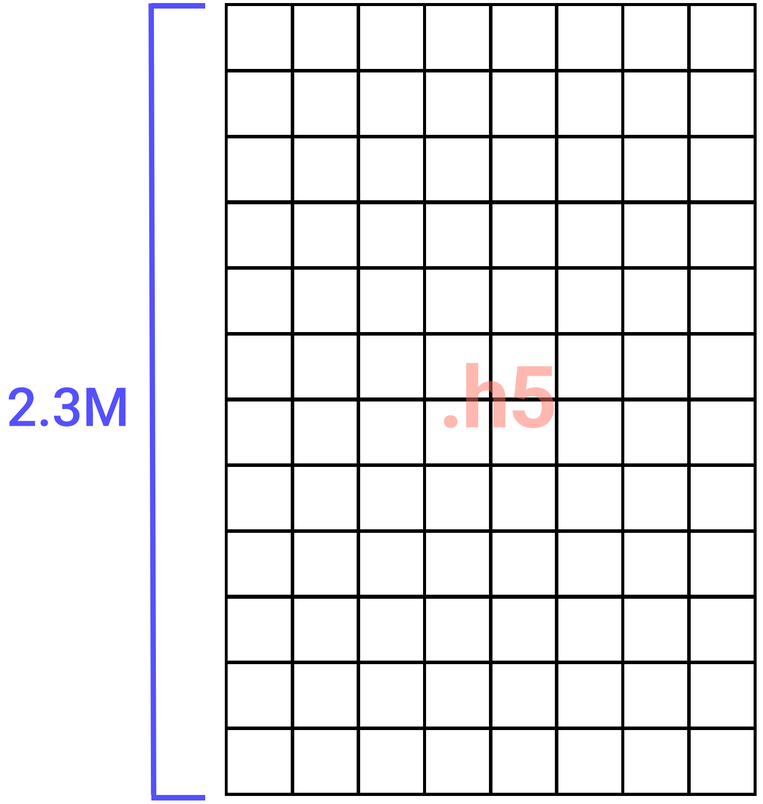


n = 10

if unique & lenght > 20



OVCAR embedding dictionary



AC embedding dictionary

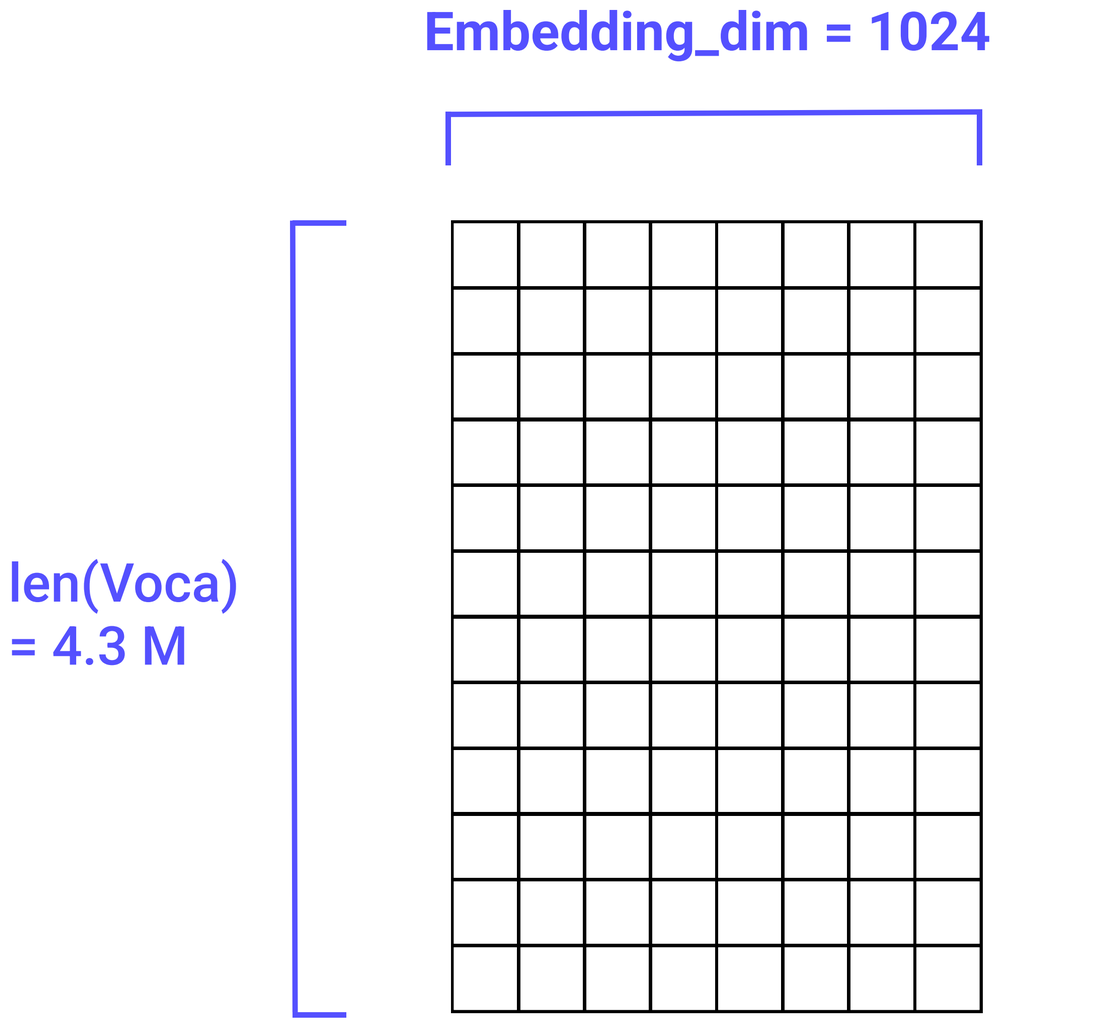
METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation



Combined embedding dictionary

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

Example:

1. Count the number of sequence included in dict in the sample

---

{
'MPYSSNFLHNRVILRVKTKT': 1,
'MGYHTSHPGHEMTANYQGIVV': 1,
'MVAERMVCPFSKSREAAEHFGG': 86,
'MTSLHNSFYSKDTSLRTPLMTP': 196,
'MDFSHIHLLCICNKESIHEIN': 2,
...
}

2. if cnt >= cutline:
   - find the embedding in the dict

'MVAERMVCPFSKSREAAEHFGG'

[0.05120769,
0.01602149,
0.11421634,
...,
-0.04321861]

dim = 1024

METHODS

Text Preprocessing

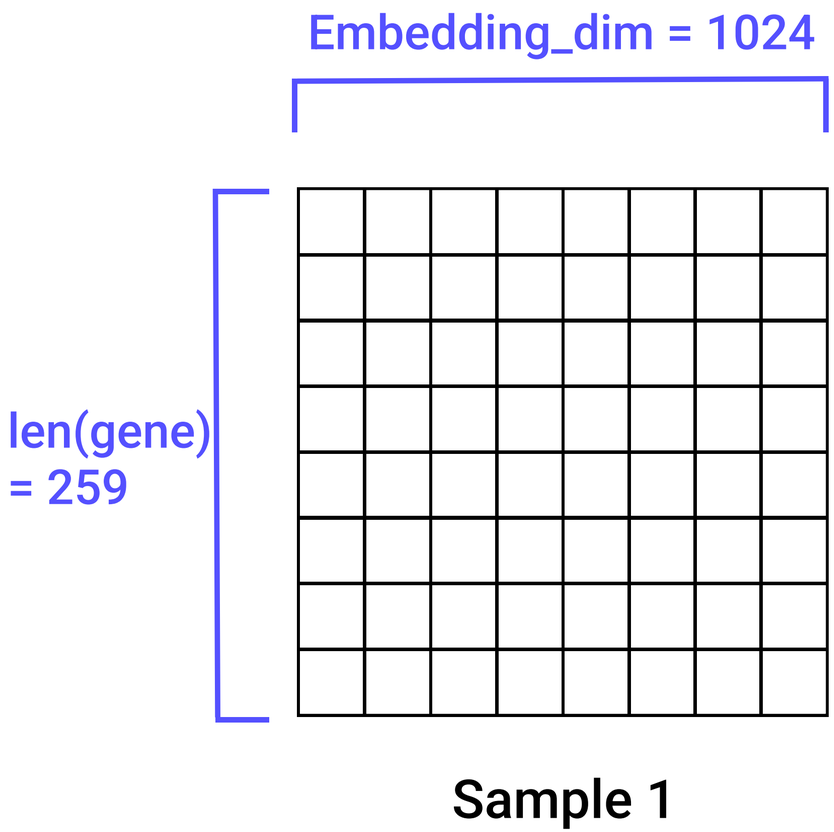
Embedding

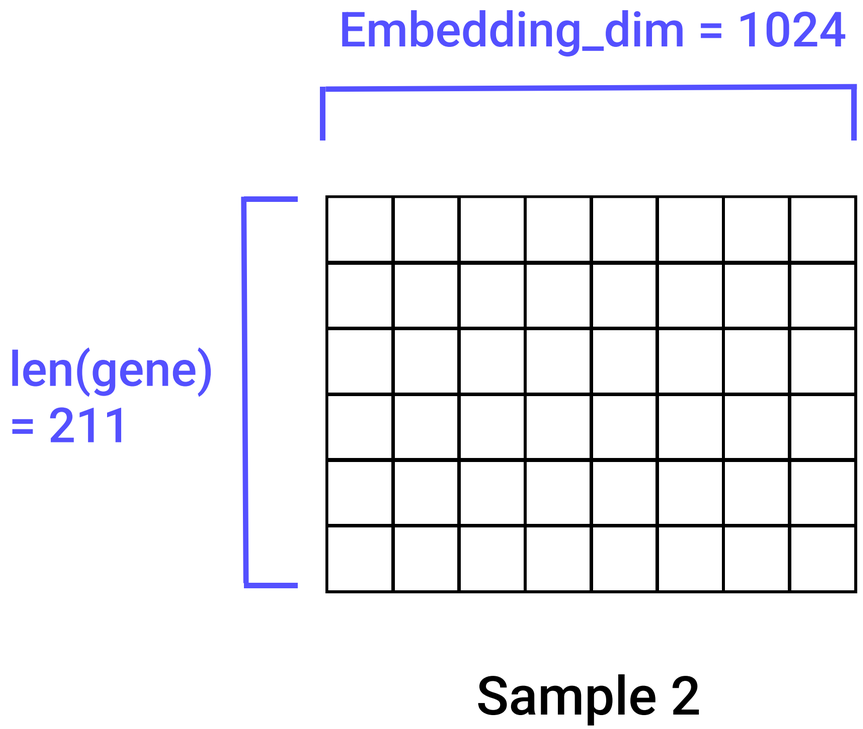
Modeling

Training & Validation

Example:

3. 2D array for each sample





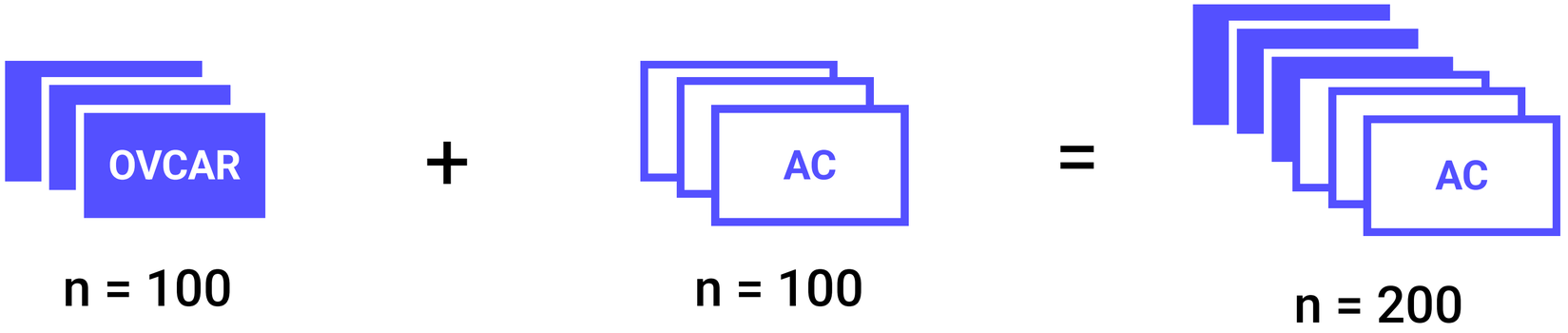
METHODS

Text Preprocessing

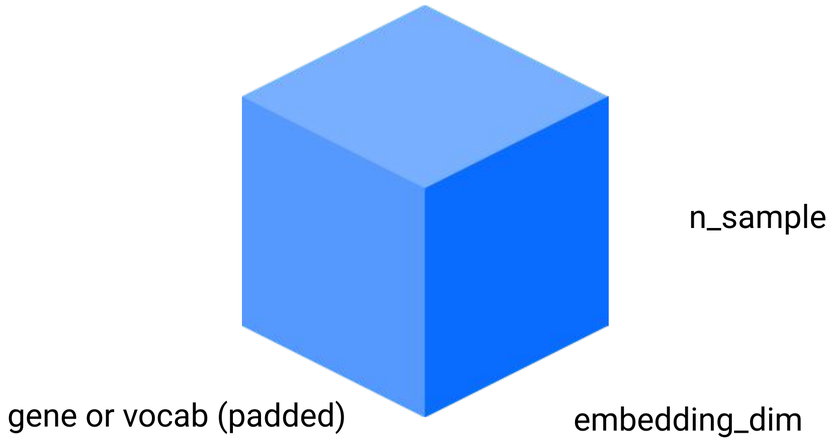
Embedding

Modeling

Training & Validation



4. Padding by mean, median, or max



Numpy array

- v2_100_max_1261.npz
- v2_100_mean_340.npz
- v2_100_median_292.npz
- v2_50_max_2461.npz
- v2_50_mean_686.npz
- v2_50_median_592.npz

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

Neural Network

	LSTM	Transformer	Multi-Head Attention	CNN
Definition	a model that handles sequential dependencies by **learning long-term** and **short-term patterns**	a model that processes sequences using **self-attention**	a model that uses **multiple attention** mechanisms	a model that applies **convolutional filters** to **extract local patterns** in data
Effective datasets type	datasets where **temporal** or **sequential order** matters	datasets where the **relationships span across** distant genes.	datasets with **diverse contextual** information	effective for data with **spatial** or structured relationships

METHODS

Text Preprocessing

Embedding

Modeling

Training & Validation

Neural Network

	LSTM	Transformer	Multi-Head Attention	CNN

Layers	• **LSTM**
	  • **2 layers**
	  • **Bidirectional**
	• Linear
		• **TransformerEncoder**
		  • **2 Layers**
		  • **8 heads**
		• Linear
			• **MultiheadAttention**
			  • **8 heads**
			• **LayerNorm**
			• **Linear**
			• **Dropout**
			• Linear
				• **Conv1D**
				• **MaxPool1D**
				• **Linear**
				• Linear

# of param	1,577,474	18,279,938	4,331,906	10,273,218

METHODS

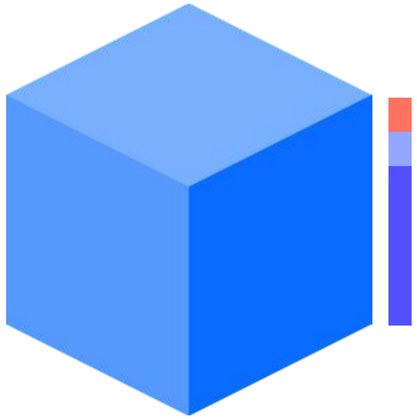
Text Preprocessing

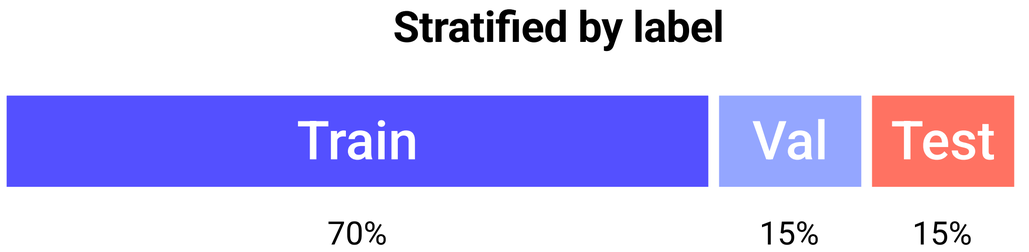
Embedding

Modeling

Training & Validation

1. Data Spliting





2. Feature Scaling by Standard Scaler

$$X_{scaled} = \frac{X - \mu}{\sigma}$$

3. NumPy to Tensor

4. Optimizer & Learning Rate Scheduler

- Adams
- ReduceLROnPlateau
- n_epoch = 100
- batch_size = 4

Summary for methods

METHODS

Text Preprocessing

Embedding

Modeling

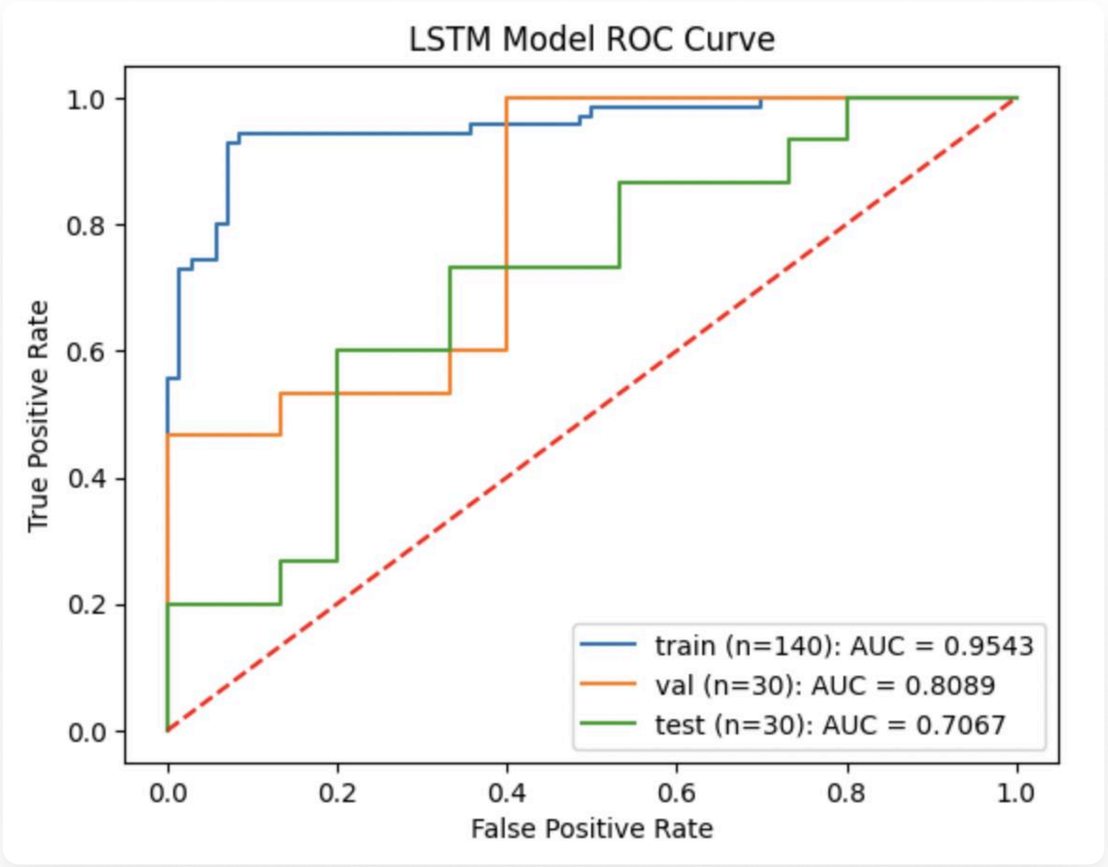
Validation

Results

Insights from Contextual Genomic Analysis

RESULT

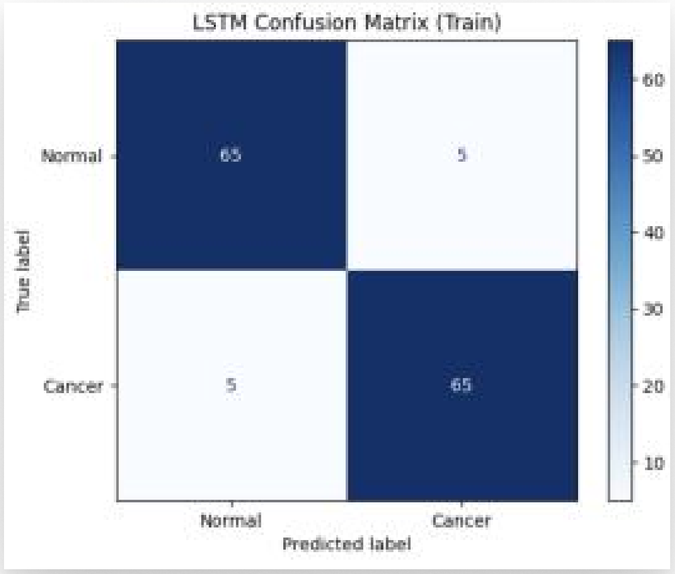
Best Model: LSTM with data/v2_50_max_2461.npz



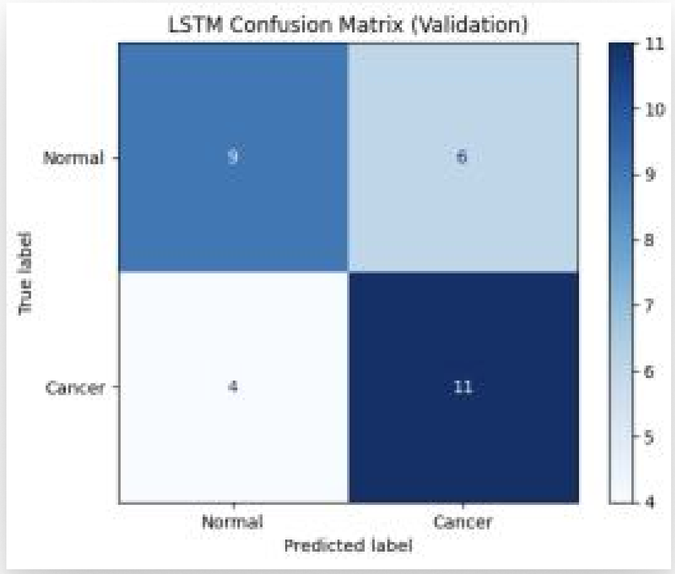
Train AUC = 0.9543
Val AUC = 0.8089
Test AUC = 0.7067

RESULT

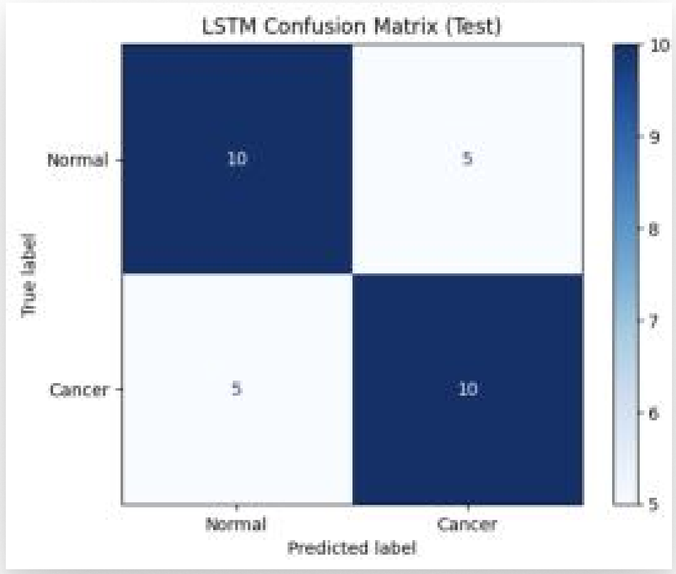
Best Model: LSTM with data/v2_50_max_2461.npz



Train



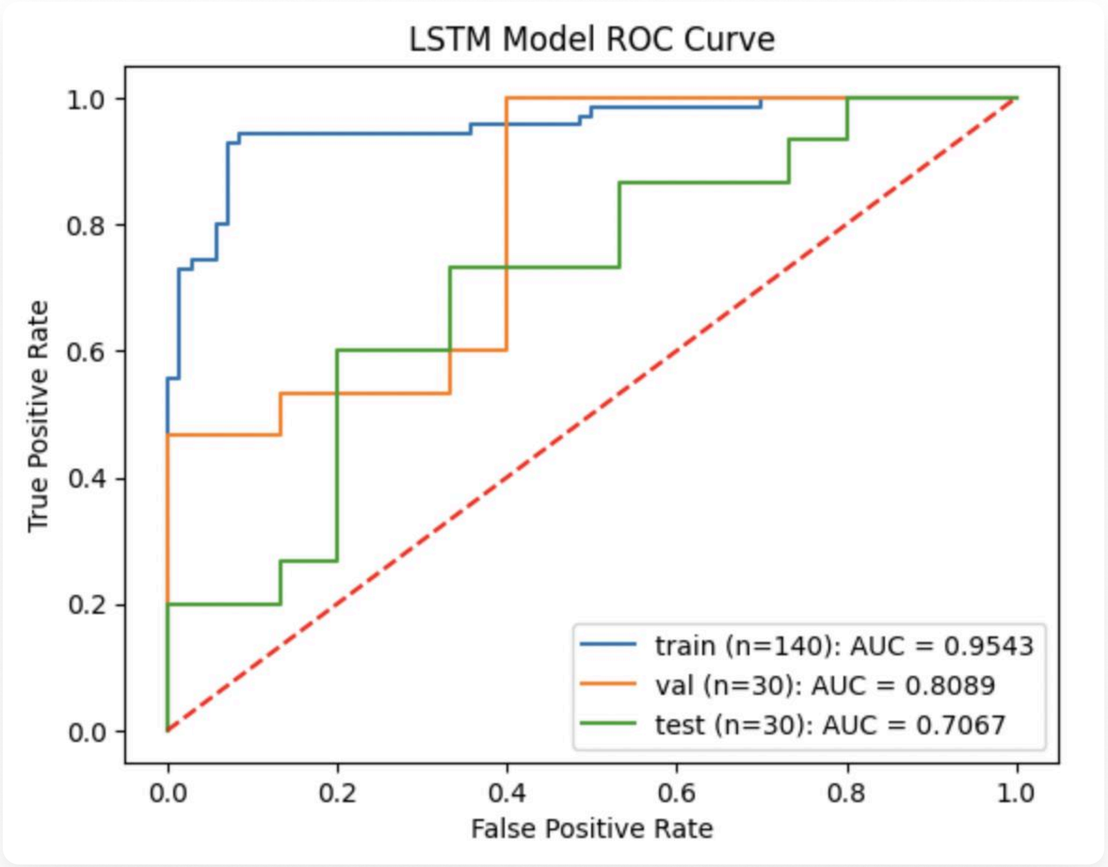
Val



Test

RESULT

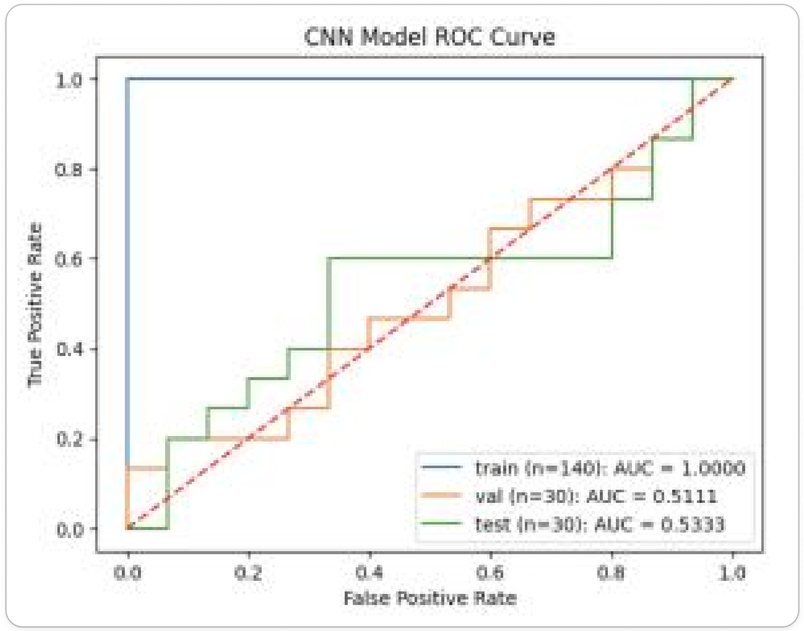
Best Model: LSTM with data/v2_50_max_2461.npz



Train AUC = 0.9543
Val AUC = 0.8089
Test AUC = 0.7067

RESULT

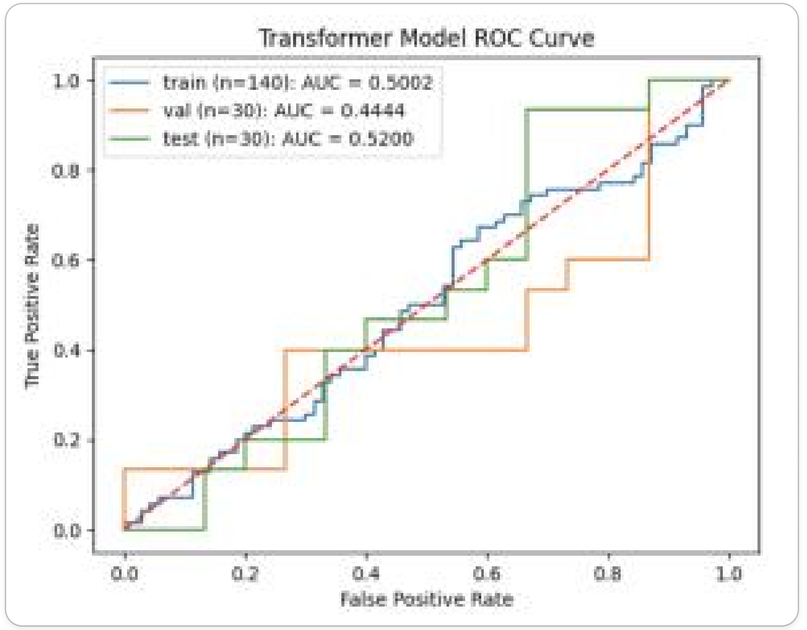
Other Model's Best Score (by test data AUC)



CNN

Val AUC = 0.5111
Test AUC = 0.5333

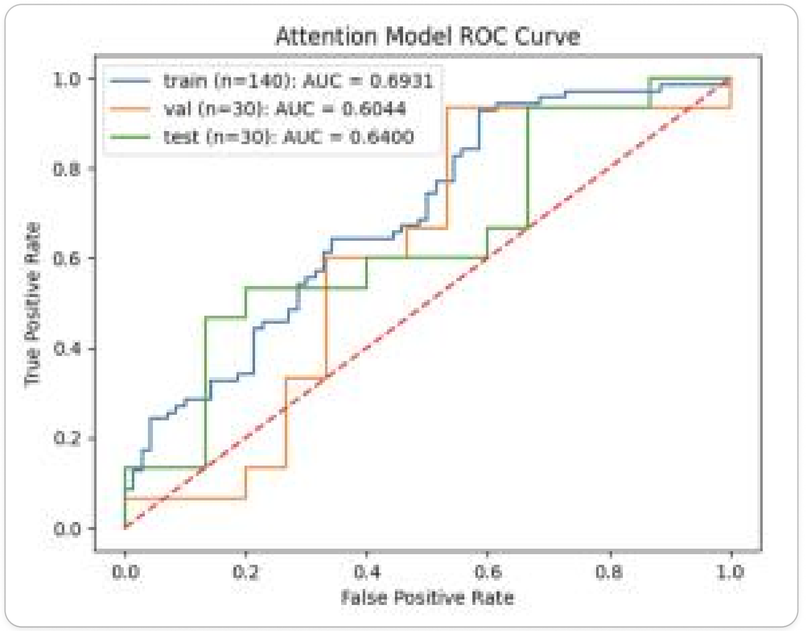
data/v2_100_max_1261.npz



Transformer

Val AUC = 0.4444
Test AUC = 0.52

data/v2_50_median_592.npz



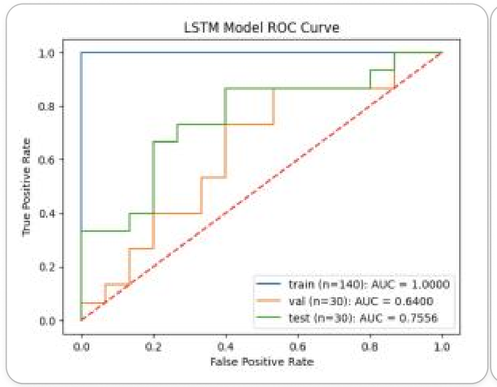
Attention

Val AUC = 0.6044
Test AUC = 0.64

data/v2_100_max_1261.npz

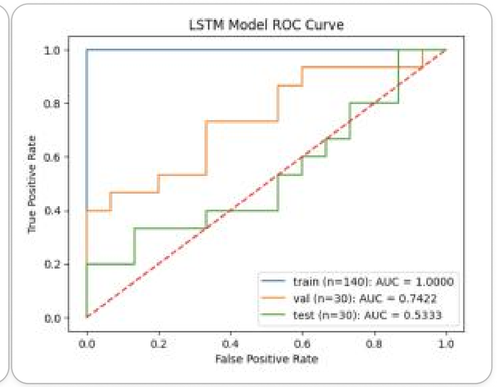
RESULT

Best Model: LSTM with other dataset



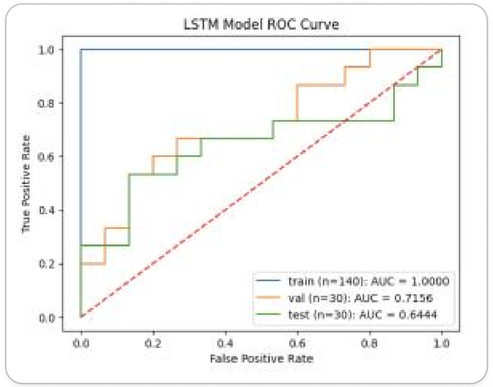
cnt_cutline = 100 padded by median (n=292)

Val AUC = 0.64
Test AUC = 0.7556



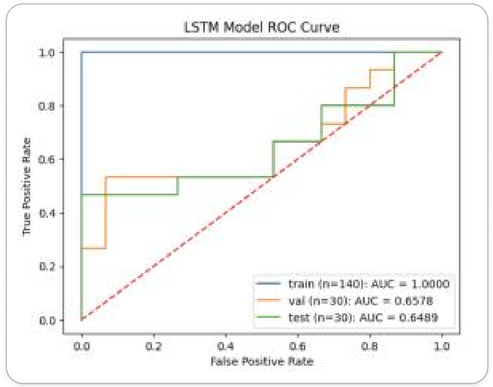
cnt_cutline = 50 padded by median (n=592)

Val AUC = 0.7422
Test AUC = 0.5333



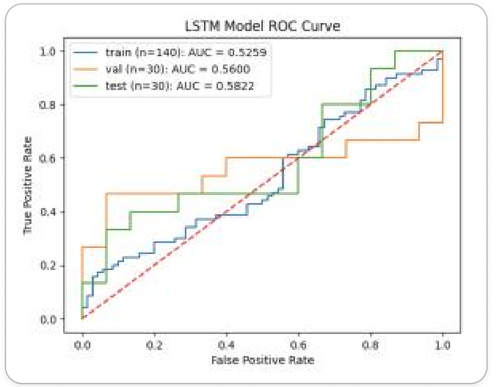
cnt_cutline = 100 padded by mean (n=340)

Val AUC = 0.7156
Test AUC = 0.6444



cnt_cutline = 50 padded by mean (n=686)

Val AUC = 0.6578
Test AUC = 0.6489



cnt_cutline = 100 padded by max (n=1261)

Val AUC = 0.56
Test AUC = 0.5822

DISCUSSION

Model Performance

The LSTM model achieved the highest performance, with a test AUC of 0.7067.

LSTM captured **sequential dependencies** in genomic data.

Potential **overfitting** issue

Comparison Across Models

Other models' test AUC values were significantly lower than LSTM.

This highlights that **sequential order** in genomic data plays a critical role.

Generalization Challenges

Variability in AUC scores across datasets

Performance is sensitive to **input characteristics,** such as embedding cutlines.

DISCUSSION

Strengths

- Simple preprocessing
- Quantitative & qualitative approaches
- LSTM’s ability underscores its suitability for genomic sequence analysis.
- Potential of applying NLP to genomic data.

Limitations

- The limited dataset size
- CNN, Transformer, and Attention models require further optimization.

CONCLUSION

Genomic sequences as language

ProtBERT for contextual insights

Promising performance in cancer diagnosis.

Thanks for your attention

If you have any question, please do not hesitate to ask.