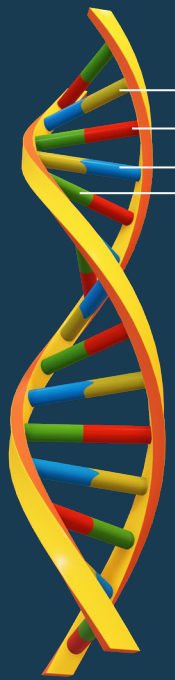
TEXAS A&M UNIVERSITY
Engineering

DNA/RNA/Single -Cell Language Models

Student: Omnia Sarhan
Instructor: Yu Zhang

DNA Vocabulary



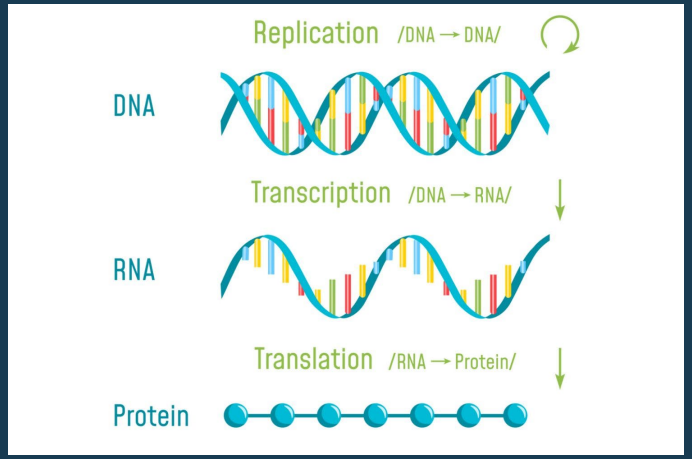
A (Adenine)

G (Guanine)

T (Thymine) (In RNA, replaced by U (Uracil))

C (Cytosine

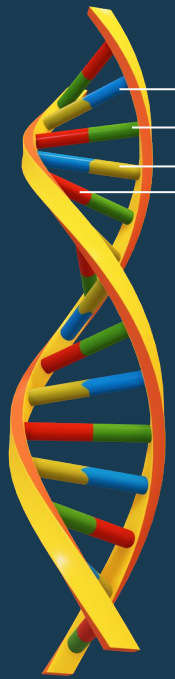
- DNA Pairs: A-T, C-G
- RNA Pairs: A-U, C-G
- Gene Sequences: ATG CCG TAA



DNA Tokens

Instead of using single letters:
k-mers (short subsequences of length k).

- k=3 (3-mer): "ATGCGT" → [ATG, TGC, GCG, CGT]
- k=6 (6-mer): "ATGCGTAC" → [ATGCGT, GCGTAC]

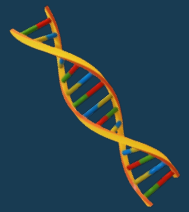


A (Adenine)
G (Guanine)
T (Thymine) (In RNA, replaced by U (Uracil))
C (Cytosine

Token Embeddings

- MASK tokens: masked during pre-training
- CLS tokens: meaning of entire sentence [whole sequence]
- SEP tokens: sentence operator/ end of sequence
- UNK tokens: Unknown
- PAD Tokens: Padding for short setences

Table of contents

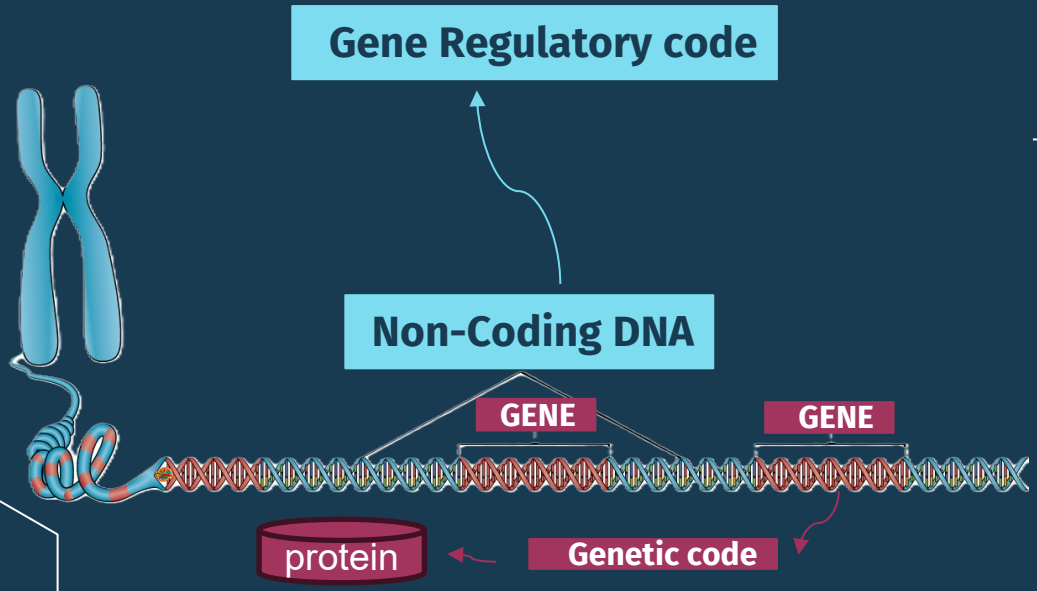


01.

DNABERT

Pre-trained Bidirectional Encoder Representations from Transformers Model for DNA-Language in Genome

Introduction



Problem Statement

- Deciphering Non-Coding DNA for hidden instructions is challenging.
- Traditional models fail to capture long-range dependencies and polysemous relationships within DNA sequences.

Objectives

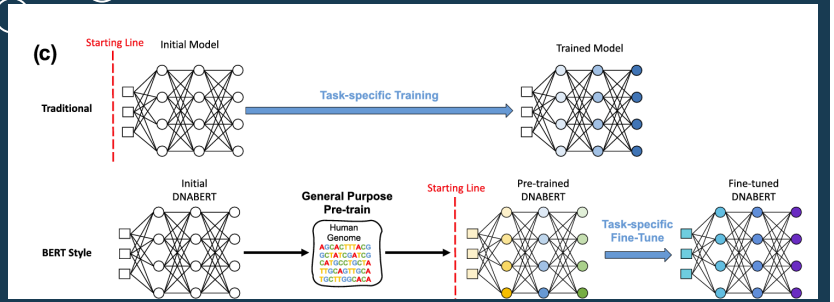
Capture global and transferable contextual information from DNA sequences

Outperform traditional deep learning models in various genomic tasks

Provide visualization mechanisms for interpretation of sequence motifs

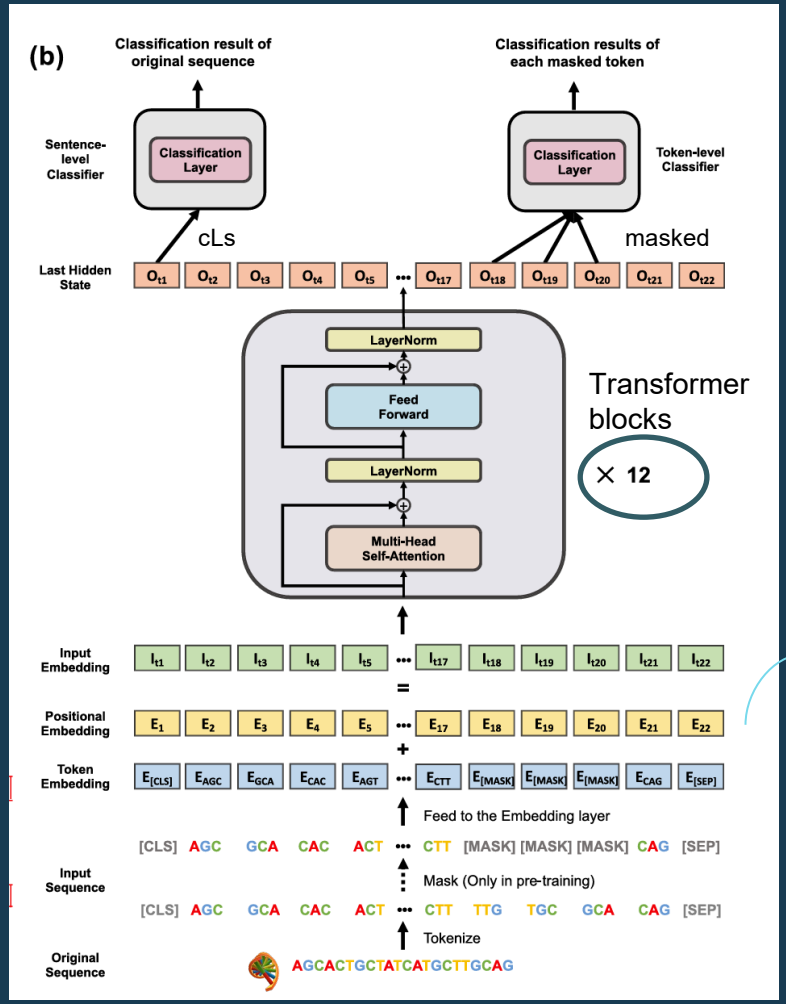
Demonstrate cross-organism applicability

Facilitate fine-tuning on task-specific datasets



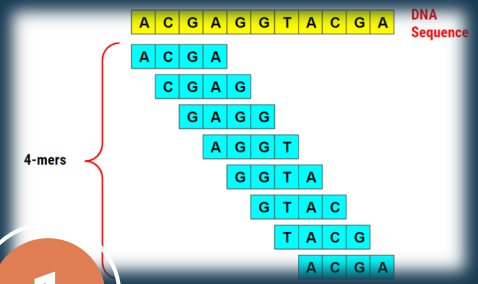
DNABERT Model

- BERT-based (same architecture)
- Attention based transformer
- Adopts pre-training +fine-tuning



Maybe sinusoidal

Methodoly



1 Tokenization

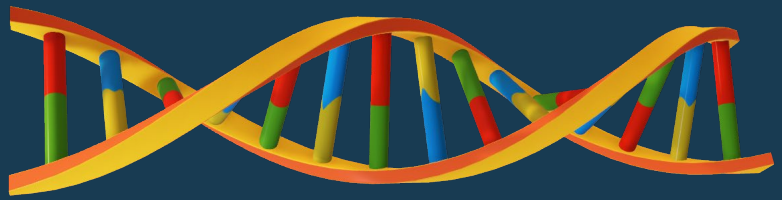
- k-mer representation instead of single nucleotides.
- Different values of k (3, 4, 5, 6)
- Added special tokens like [CLS], [PAD], [UNK], [SEP], and [MASK]

2 Pre-training

- masked language modeling (MLM) for random masking [15%]
- Human genome (5-510 base pairs)
- 12 Transformer layers, 768 hidden units, and 12 attention heads

3 Fine-tuning

- Task-specific datasets
- Long sequences exceeding 512 tokens are split and processed as DNABERT-XL.
- Best = DNABERT-6
- Skip masking



Results

DNABERT

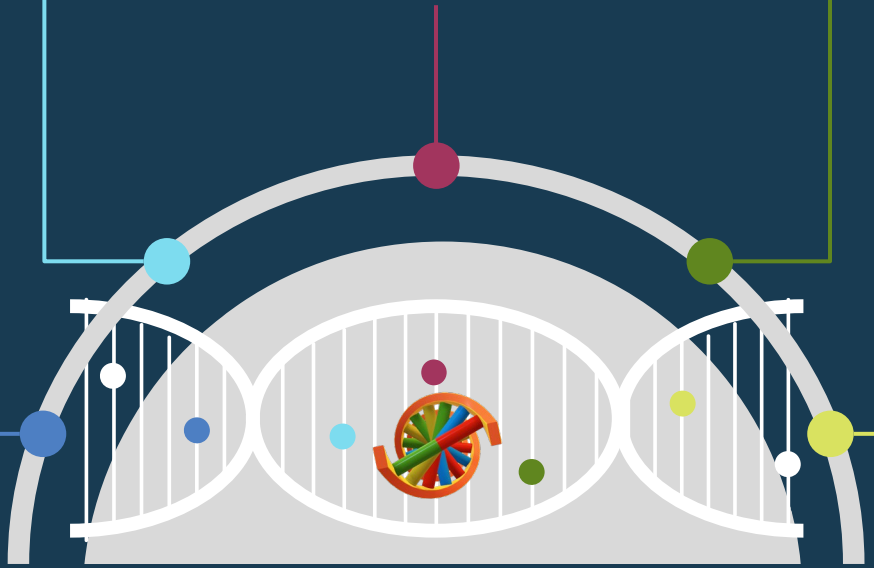
- Generalize over tasks
- Identifying functional genetic variants

DNABERT-Splice

Accurately recognizes canonical and non-canonical splice sites

DNABERT-viz

Allows visualization of important regions, contexts and sequence motifs



DNABERT-TF

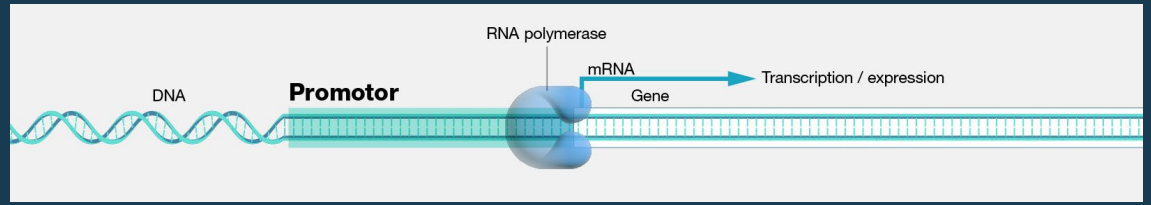
Accurately identifies transcription factor binding sites

DNABERT-Prom

Effectively predicts proximal and core promoter regions

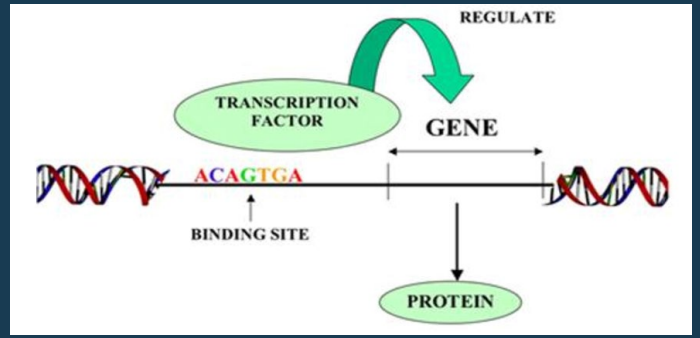
Applications

DNABERT-Prom



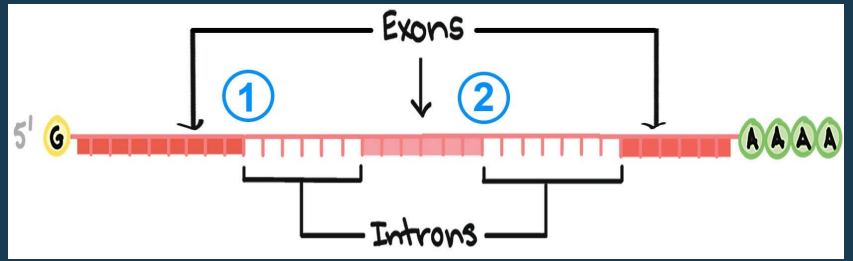
DNA sequence that initiates the transcription of a gene

DNABERT-TF



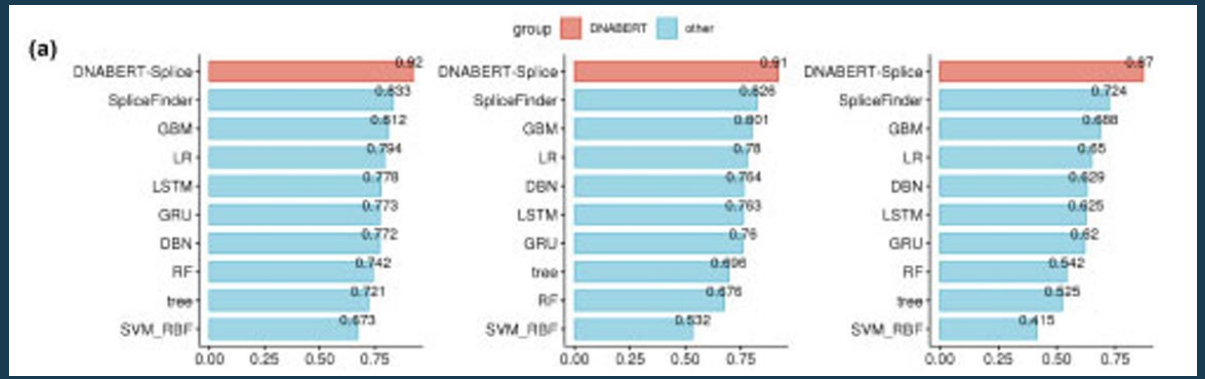
functional genetic variants

DNABERT-Splice

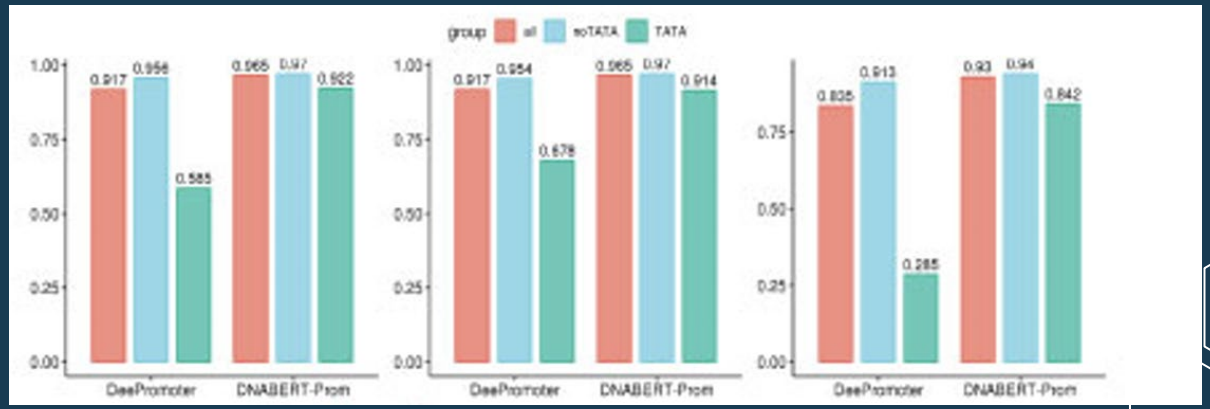


Results: (left to right) accuracy, F 1 and MCC

Splice

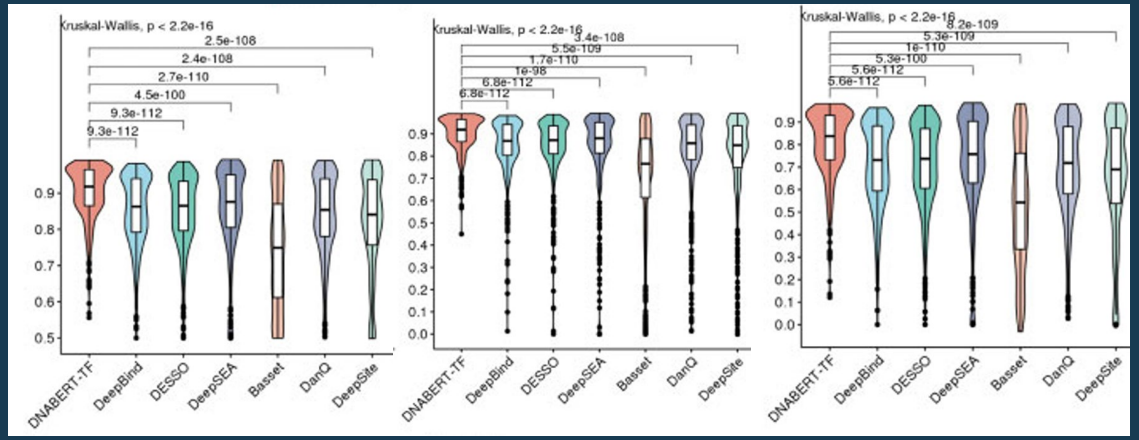


Prom

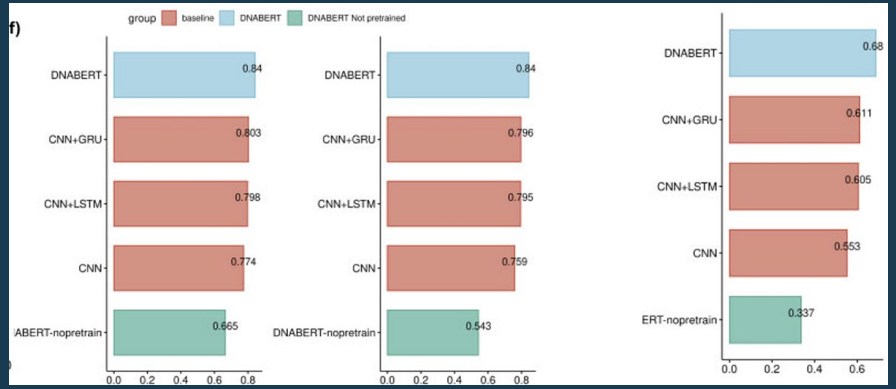


Results: (left to right) accuracy, F1 and MCC

TF



General
(mouse
encode)



Future Work

1. Other sequence prediction tasks

Determining CREs and enhancer regions from ATAC-seq and DAP-seq

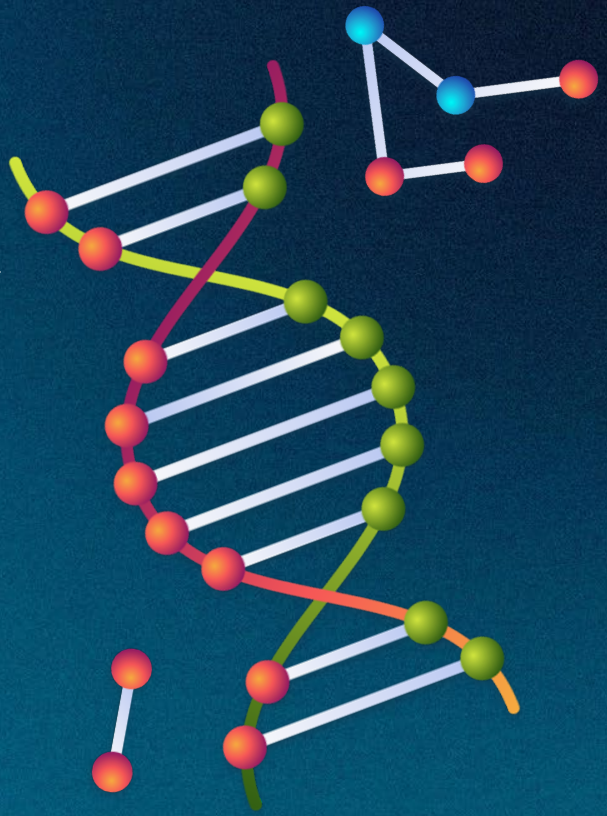
2. Prediction of binding preferences of RNA-binding proteins (RBPs)

3. Direct machine translation on DNA

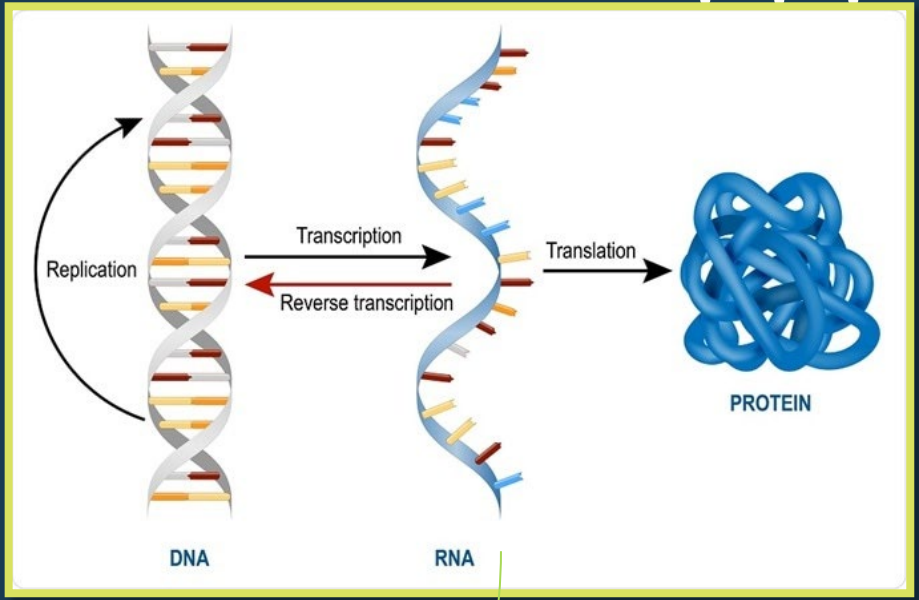
02.

5' UTR

A Language Model for Decoding Untranslated Regions of mRNA and Function Predictions



Background


Replication
Transcription
Reverse transcription
Translation
PROTEIN
DNA
RNA


5′ Cap - plays a critical role in translational yield and nucleic acid stability *in vivo*
Poly(A) tail - protects the mRNA from nuclease degradation
5′ Cap
5′ UTR
Open reading frame
3′ UTR
Poly(A)
5′ untranslated region (UTR)
5′ UTR - regulates protein expression levels and translation initiation
3′ UTR - regulates protein expression levels and influences the stability and half-life of the mRNA

Introduction

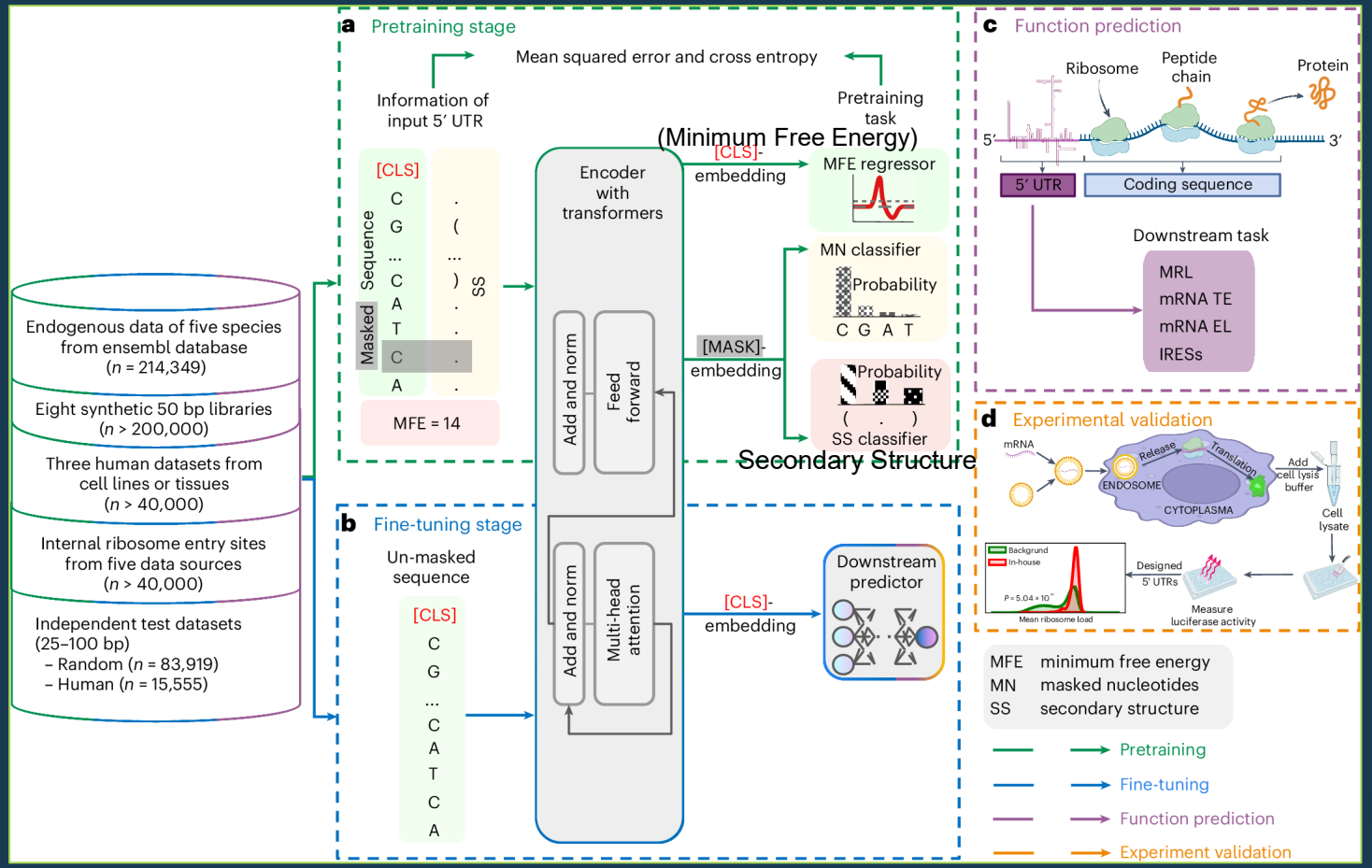
Problem Statement

No unified foundation model to study function of 5'UTR

Objectives

Use Language model to Extract meaningful semantic representations from UTRs of raw mRNA sequences and map them to predict functions of interest.

5'UTR-LM Model Overview



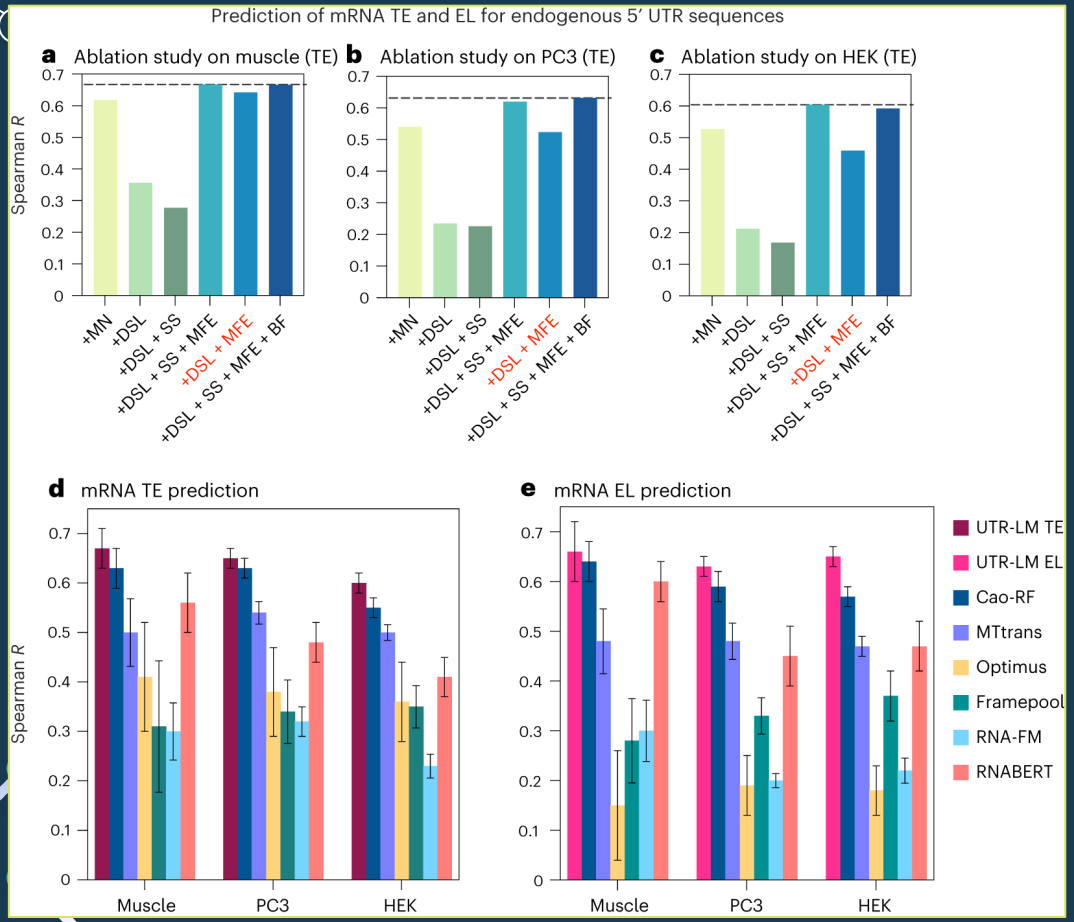
Results

UTR-LM predicts the ribosome loading

URR-LM identifies IRESs

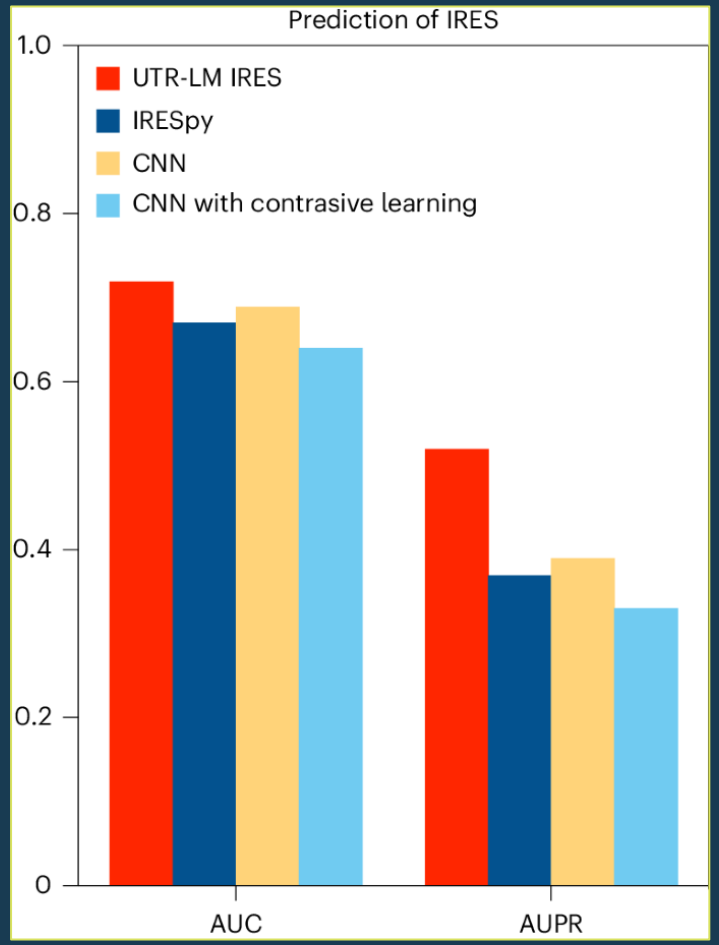
UTR-LM predicts mRNA TE and expression

New designs validated in wet-lab experiments



UTR-LM predicts mRNA TE and EL

URR-LM identifies IRESs



Conclusion

Conclusion

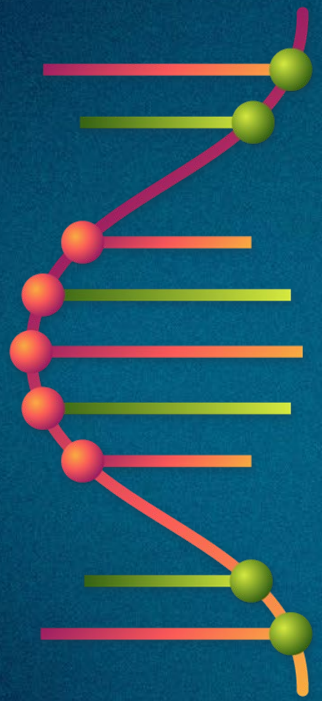
- Outperforms the best-known baseline in each task.
- Performance not limited by sequence length

limitations

- Computationally expensive

Future

sparse transformers for modelling longer RNA sequences and more complex biological functions

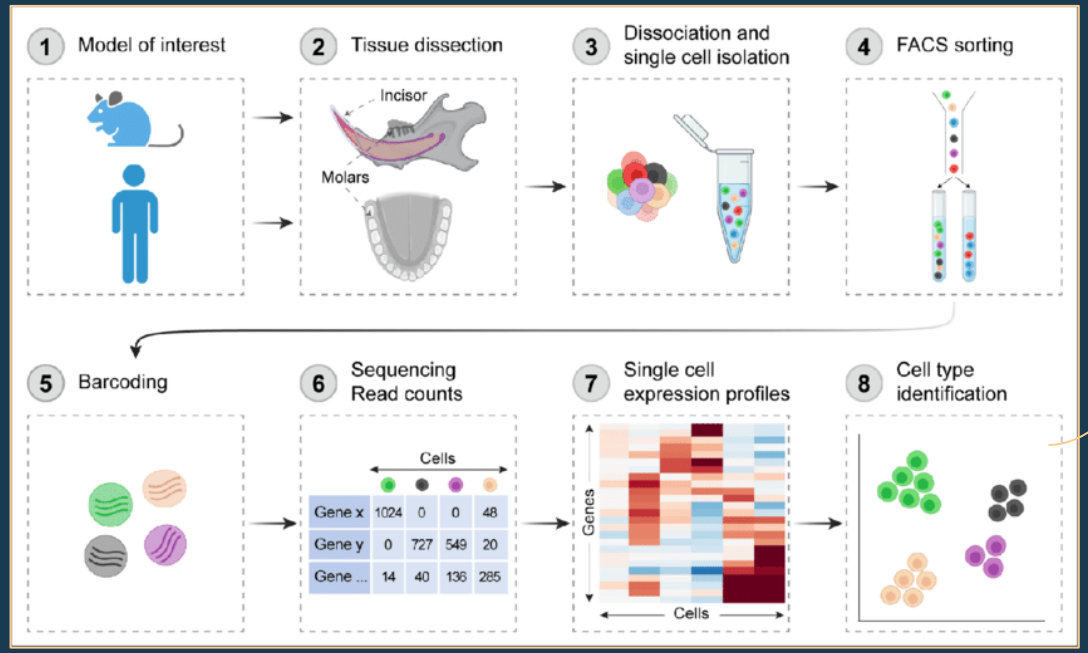


03.

scGPT

Towards Building a Foundation Model for Single-Cell Multi-omics using Generative AI

Single -cell RNA sequencing (scRNA-seq)



personalized therapeutic strategies

cellular heterogeneity exploration

lineage tracking

pathogenic mechanism elucidation

Introduction

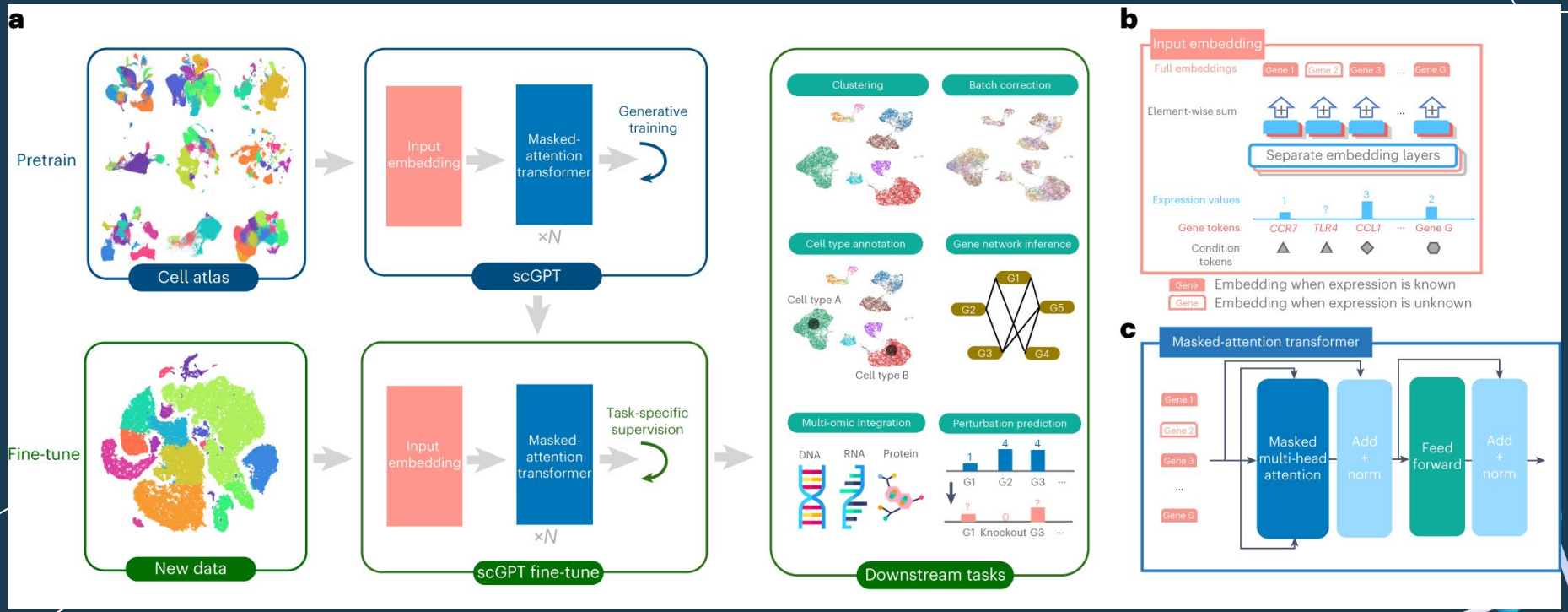
Problem Statement

Current machine-learning-based methods in single-cell research are scattered

Objectives

- Foundation model pretrained on large-scale data
- comprehend the complex interactions between genes across diverse tissues.

scGPT Model Overview



Results

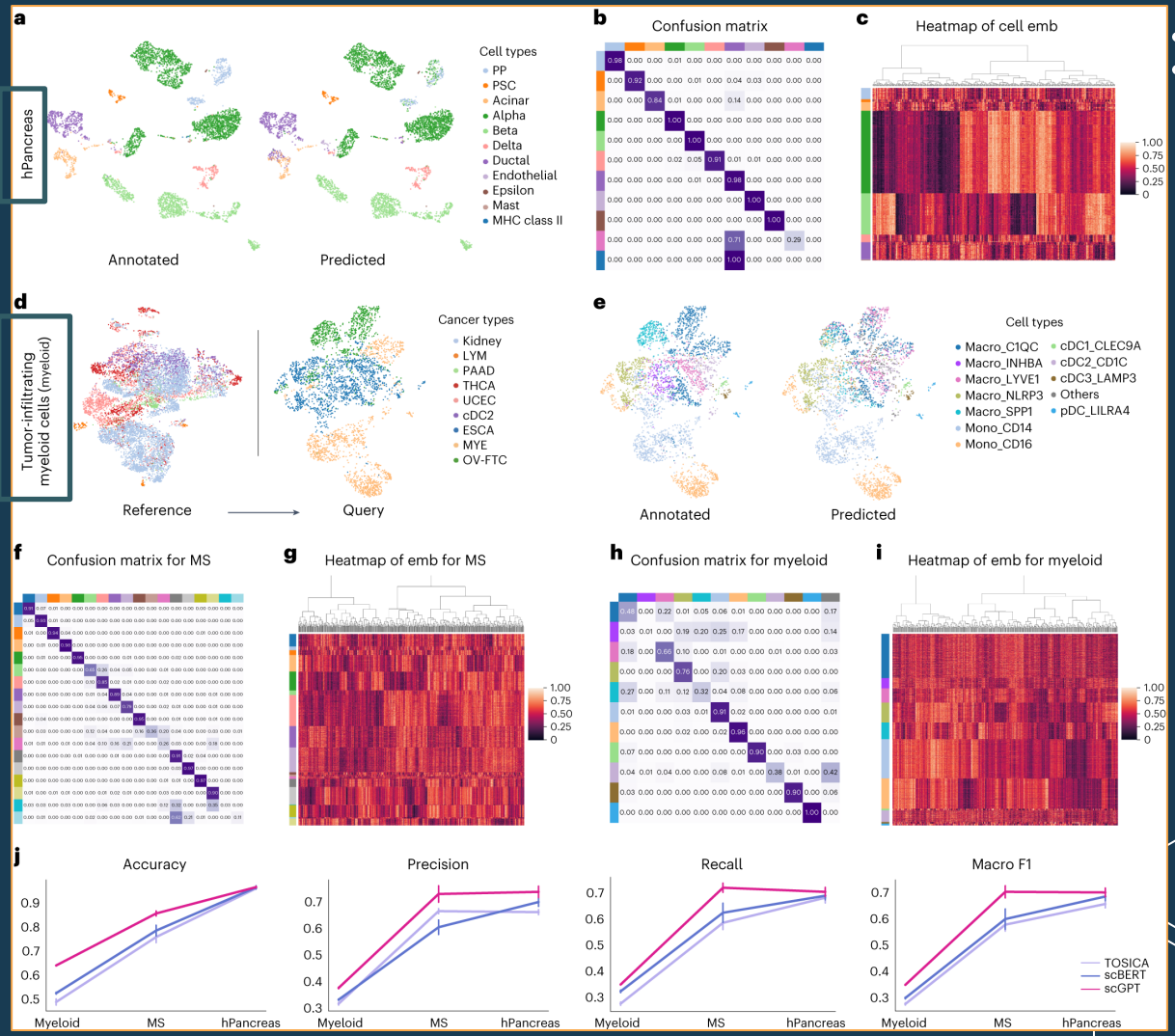
Improves the precision of cell type annotation

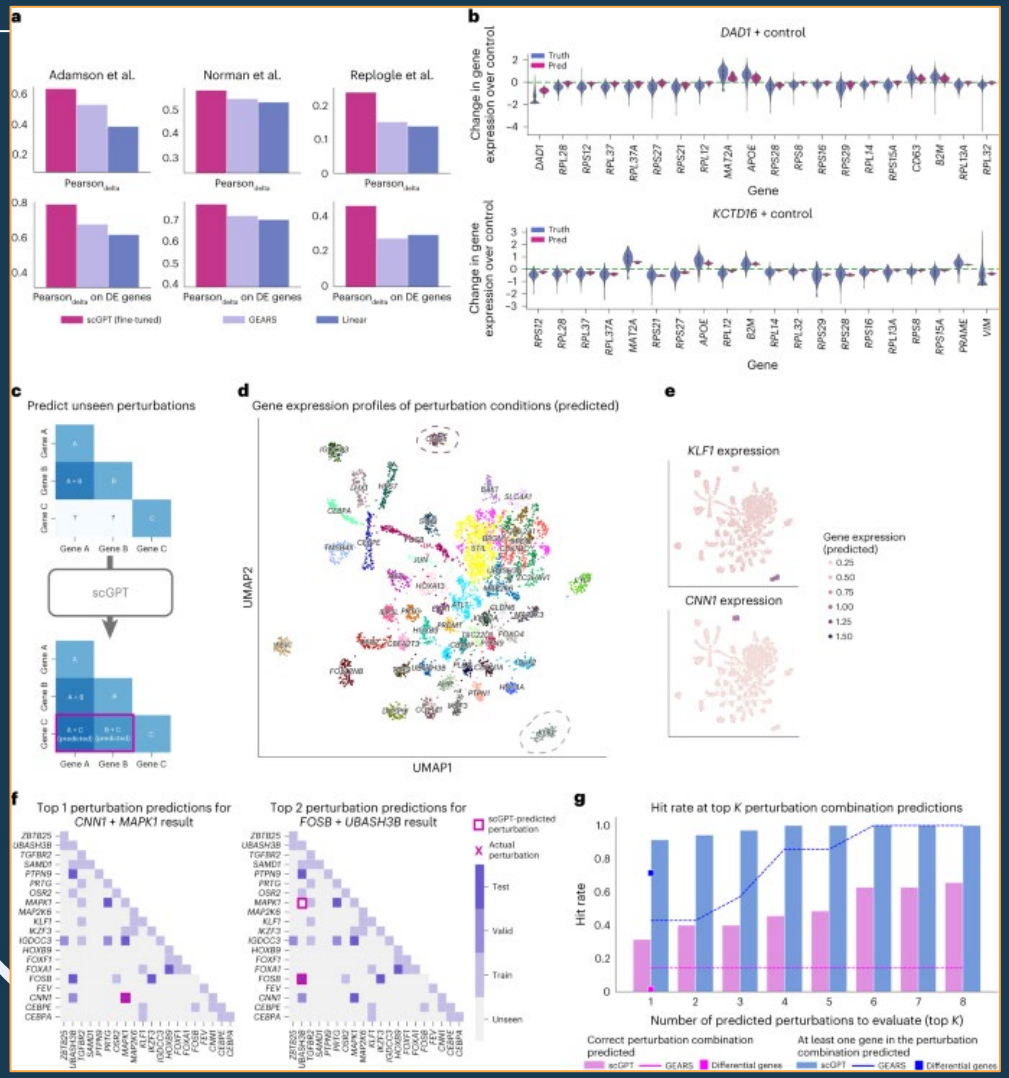
multi-batch and multi-omic integration

Predicting Unseen Genetic Perturbation Responses

Uncovers gene networks for specific cell states

Cell Type Annotation



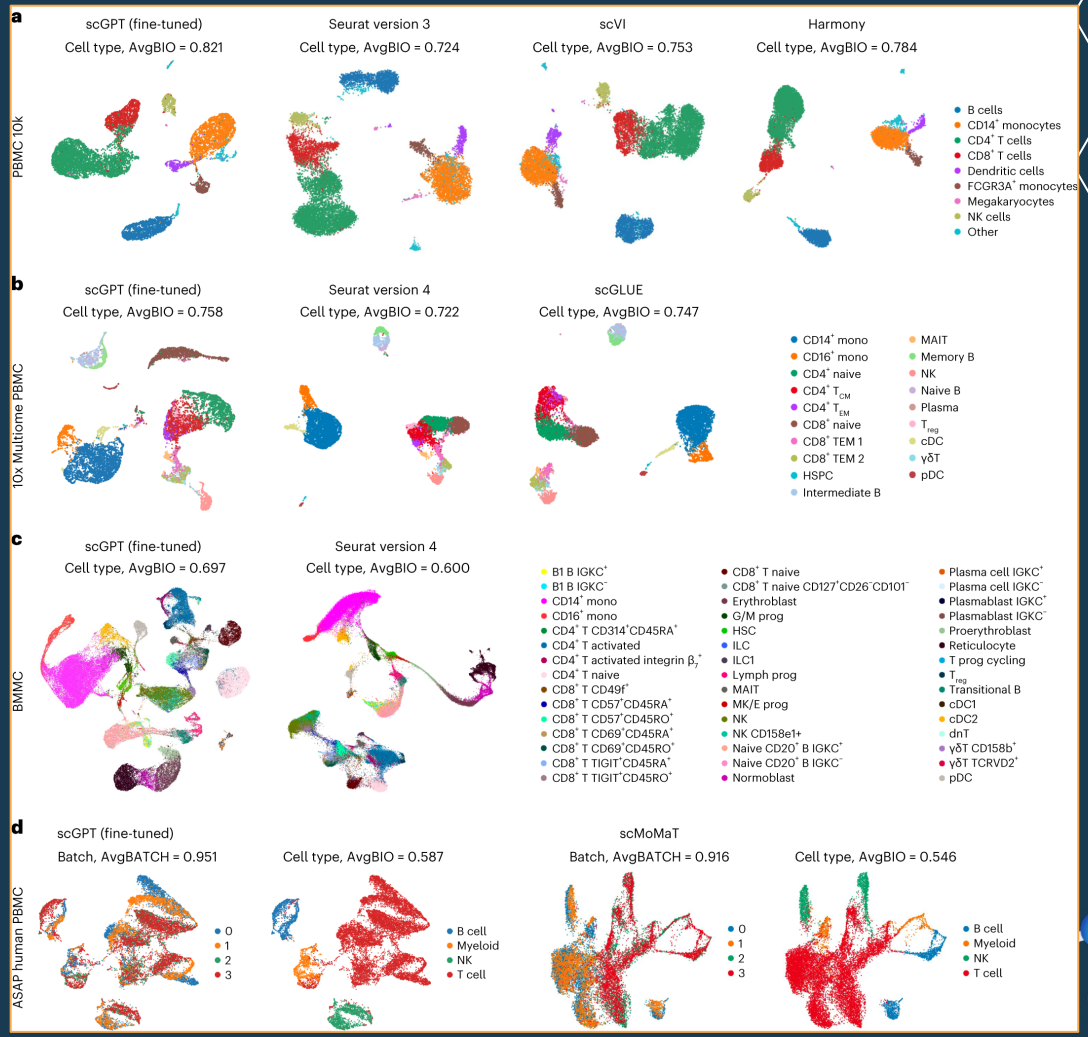


Predicting Unseen Genetic Perturbation Responses

modifications in gene expression or function caused by:

- Gene knockouts (KO) → Removing a gene entirely.
- Gene knockdowns (KD) → Reducing a gene's expression.
- Overexpression (OE) → Increasing gene activity.

Multi -Batch & Multi -Omic Integration



Conclusion

Limitations

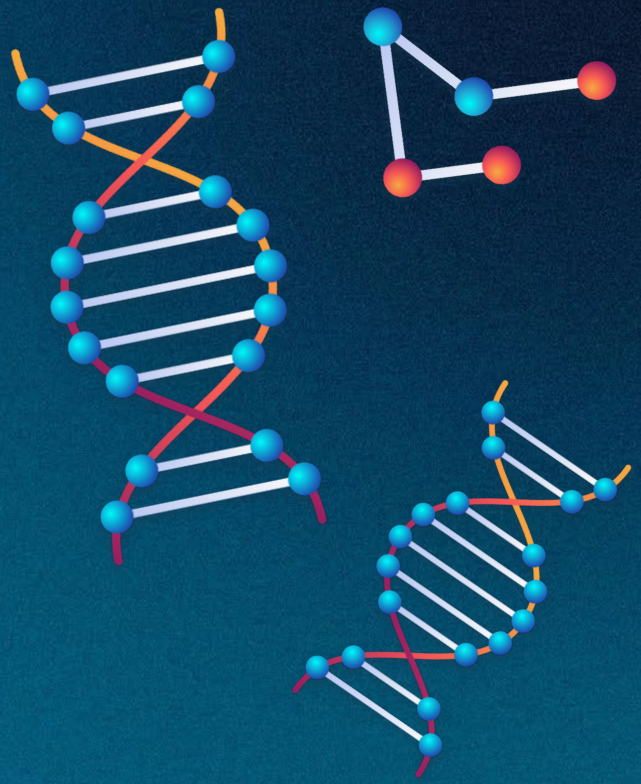
- Pretraining does not mitigate batch effects.
- zero-shot performance could be constrained on datasets with technical variation
- Evaluating the model is also complex due to variation in data quality

Future Work

- pretrain on a larger-scale dataset with more diversity
- explore in-context instruction learning for single-cell data.

Summary

Conclusion & Questions



Model	DNABERT (BERT based)	UTR-LM	scGPT
Domain	DNA sequencing	5'UTR of mRNA	(scRNA-seq)
motive	Deciphering DNA sequences	Unified foundation model to study function of 5'UTR	Unified foundation model to study single-cell RNA functions
Method	• BERT architecture • Tokenization with k-mer (6) • Modify pre-training process • Fine-tuned on 3 specific tasks • Benchmark with current tools	• Transformer-based architecture • Masked nucleotide (MN) prediction • secondary structure (SS) • minimum free energy (MFE) • Fine-tuned on multiple downstream tasks	• Transformer-based architecture • Pretrained on a large corpus of single-cell RNA data • tokenization of gene expression profiles. • Multi-task learning approach
Results	• surpassing existing tools • Enhanced performance with limited data • No- separate training needed • Flexible learning of DNA in different situations	• outperforms the best-known baseline in each task. • Performance not limited by sequence length • Validated through wet-laboratory experiments • Zero shot generalization	• Pretrained model extrapolates to unseen datasets. • Outperform existing models • High accuracy in cell type annotation • strong scaling properties
limits	• Sequence Length Limitation • Dependence on k-mer Tokenization	• Computationally expensive	• Pretraining does not mitigate batch effects. • zero-shot performance could be constrained on datasets with technical variation

Questions

Paper	Question
ALL	It appears that these three papers directly apply LLMs to gene sequence inputs. Are there any studies that explore incorporating a separate encoder for processing the gene sequence, enabling the model to handle multimodal inputs (text + gene data)?
DNABERT	Do the authors mention why they stop at k=6 for the k-mer tokenization? Do you believe that larger k's could lead to better performance since each token might be able to capture richer context?

Any studies that explore incorporating a separate encoder for processing the gene sequence, enabling the model to handle multimodal inputs (text + gene data)?

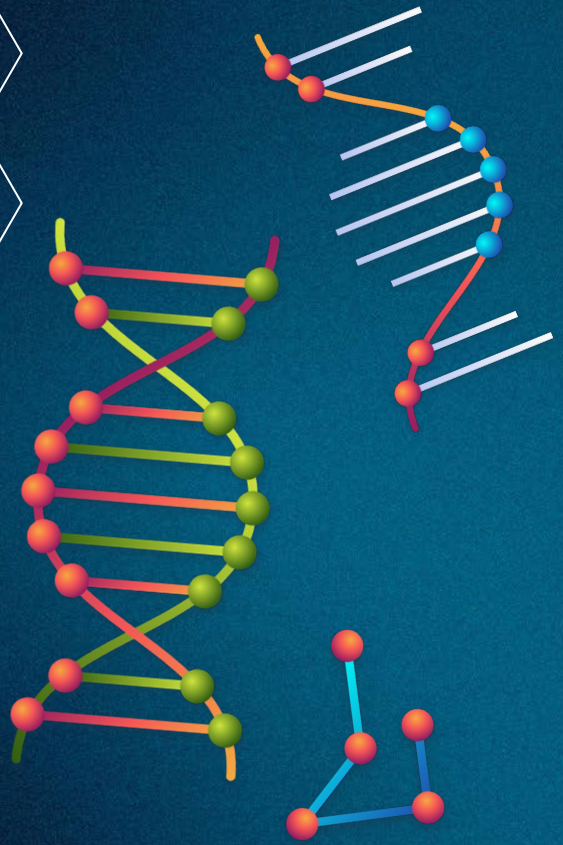
- **Multi-modal Transfer Learning Between Biological Foundation Models**
  - Uses separate encoders for DNA, RNA, and proteins, each trained independently.
  - Aggregation layers fuse embeddings from different modalities.
  - Applied for predicting RNA transcript isoforms and cross-modality generalization.

- **Prot2Text: Multimodal Protein Function Generation with GNNs & Transformers**
  - GNN encoder for protein structural data + Transformer encoder for text-based annotations.
  - Output: rich functional descriptions of proteins.
  - Beyond simple classification, enhancing explainability in protein research.

- **Geneverse: Open-Source Multimodal LLMs for Genomics & Proteomics**
  - Integrates genomic, proteomic, and textual data using specialized encoders.
  - Fine-tuned LLMs generate gene function descriptions & protein function predictions.
  - Supports tasks like spatial transcriptomics & marker gene selection.

DNABERT stops at k=6 for the k-mer tokenization? Do you believe that larger k's could lead to better performance since each token might be able to capture richer context?

- Simple Answer: NO
- ⊠ k (e.g., k=7) = ⊠ vocabulary to 16,385 tokens= ⊠ complexity & computational cost
- ⊠ k = over-specialize the model = can't generalize [overfitting]
- DNABERT-3, 4, 5, and 6 achieved very similar performance, with k=6 slightly outperforming the others = **not be significant enough to justify increase**.



Thank you for listening

Any More Questions?