Benchmarking DNA Foundation Models for Genomic Sequence Classification

Chong Wu

Department of Biostatistics

The University of Texas MD Anderson Cancer Center

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

ENAR 2025

Mar 25, 2025

Scientific questions

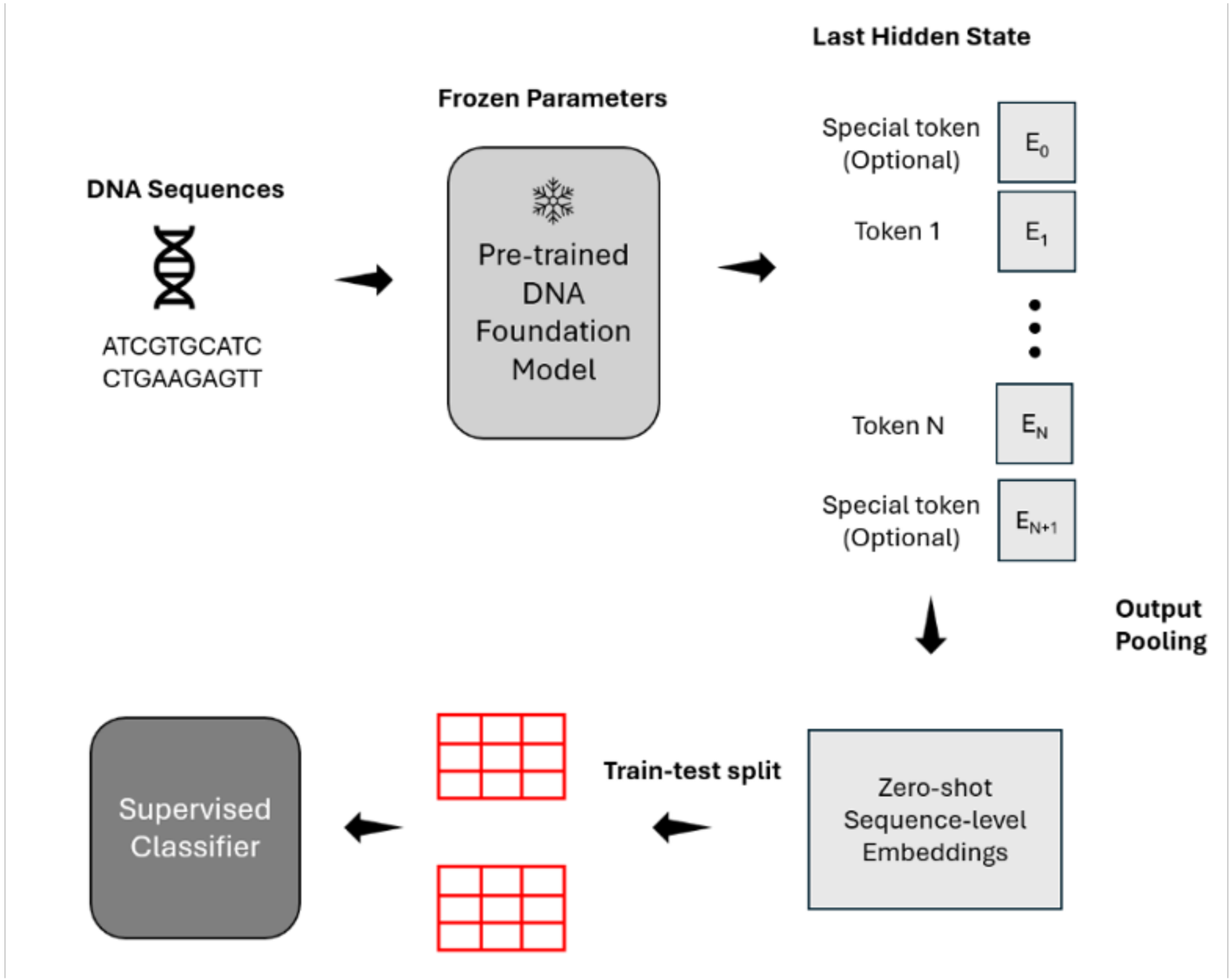
- Many DNA foundation models are currently available. How to choose the optimal DNA foundation models for subsequent analysis
- Factors to consider when using DNA foundation models:
  - Output pooling methods (mean pooling, summary-token CLS pooling, etc.)
  - Sequence length (performance on long vs short sequences)
  - Task type (binary classification, multi-class classification, regression)
  - Underlying questions (human genome? multi-species involved? epigenetic and transcriptomics involved?)
  - Comparison to baseline models (CNN, Enformer)

Benchmarking the DNA foundation models using zero-shot learning with diverse genomics tasks

Benchmarked Models

- BERT-based models: DNABERT-2, Nucleotide Transformer v2, GROVER
- Hyena-based models: HyenaDNA
- Mamba-based models: Caduceus-Ph

Key: zero-shot (other than fine-tune)



Evaluation of DNA Foundation Models in Genomic Tasks

Datasets

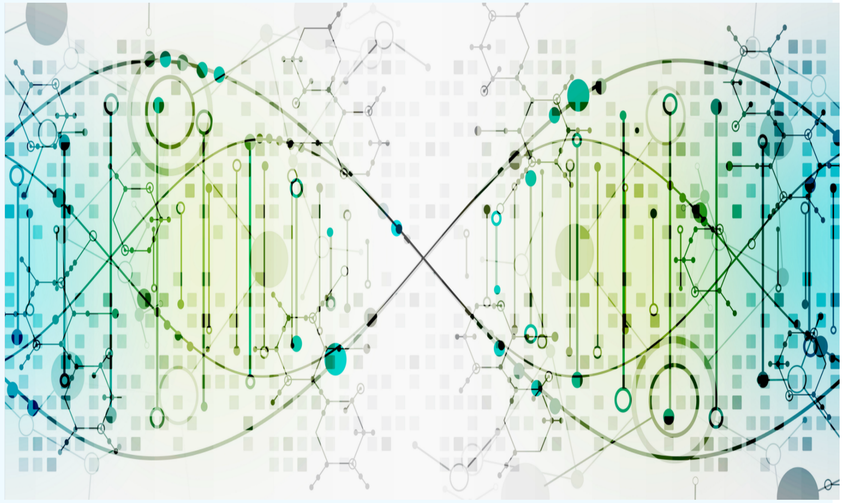
We used labelled datasets taken from the validation datasets of these DNA foundation models' original works, including

- Promoter region identification
- Transcription factor binding site identification
- Open chromatin region identification
- Splice site identification
- Covid variants classification

And also labelled datasets taken from public sources including

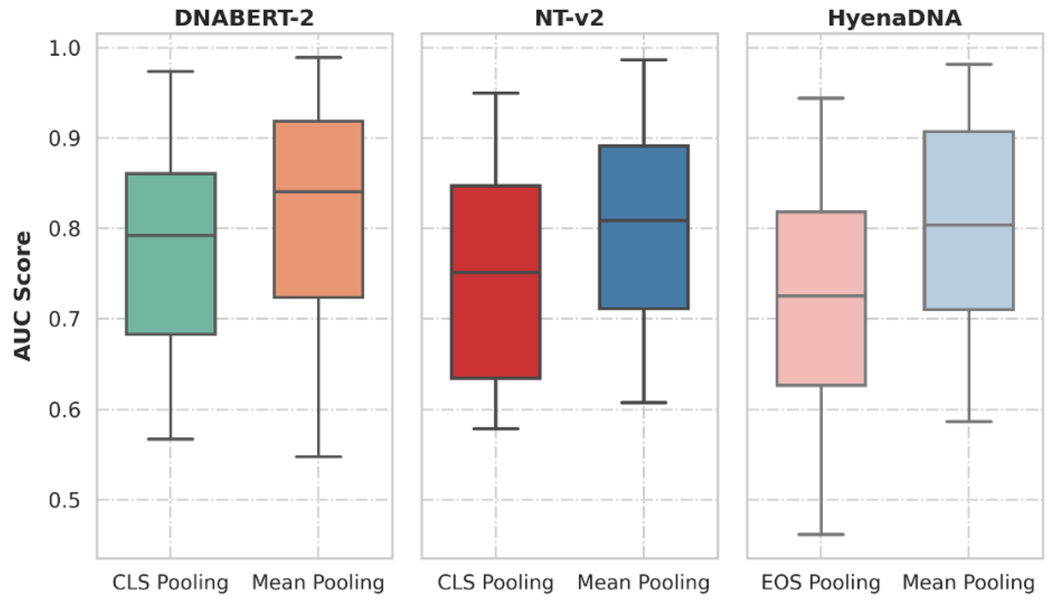
- Promoter region identification for different species
- Epigenetic modification (5mC, 6mA, 4mC) identification
- Dnase I identification

Makes a total of 57 datasets/tasks



This figure is downloaded from Google Image

Mean pooling is better



Generally, all models receive best AUCs (if # class > 2, use one-vs-rest AUC) when using output **mean pooling**, compared with summary-token pooling and maximum pooling.

Out of 57 datasets, mean pooling achieves the highest AUC in

- DNABERT-2: 50 datasets
- HyenaDNA: 43 datasets
- Nucleotide Transformer v2: 52 datasets
- GROVER: 47 datasets
- Caduceus-Ph: 49 datasets

- We also performed DeLong test to confirm that mean pooling is significantly higher for all models.

Benchmarking with zero-shot embeddings

- Now we compare across different DNA foundation models. For simplicity, we only show the results using **mean pooling** for every model here, as we previously showcased that mean pooling is simply “best” for all models.

- For better statistical illustration, we apply DeLong’s test to compare AUCs, and report the models that has **significantly higher AUC than at least 2 other models** ($p < 0.01$).

There is no uniformly best model

Human genome based region classification tasks

Dataset	Models (Order does NOT indicate anything)
Promoter_GM12878	Caduceus-Ph
Promoter_HUVEC	DNABERT-2, Caduceus-Ph
Promoter_Hela-S3	DNABERT-2, Caduceus-Ph
Promoter_NHEK	Caduceus-Ph, GROVER, DNABERT-2
promoter_notata_251bps	Caduceus-Ph, GROVER
promoter_notata_70bps	Caduceus-Ph, NT-v2, GROVER
promoter_tata_70bps	Caduceus-Ph, NT-v2
promoter_all_70bps	Caduceus-Ph, NT-v2, GROVER
promoter_notata_300bps	Caduceus-Ph
promoter_tata_300bps	HyenaDNA
promoter_all_300bps	Caduceus-Ph, NT-v2
coding	Caduceus-Ph, GROVER, DNABERT-2
donors	DNABERT-2, Caduceus-Ph, GROVER
acceptors	DNABERT-2, Caduceus-Ph, GROVER
enhancer	DNABERT-2, NT-v2
enhancer_cohn	DNABERT-2, Caduceus-Ph, GROVER
enhancer_ensembl	Caduceus-Ph, NT-v2, GROVER
Human_TFBS_1	Caduceus-Ph, GROVER
Human_TFBS_2	Caduceus-Ph, GROVER
Human_TFBS_3	Caduceus-Ph
Human_TFBS_4	Caduceus-Ph, GROVER
Human_TFBS_5	Caduceus-Ph, GROVER
open_chromatin_region	Caduceus-Ph, GROVER, DNABERT-2

Multi-species genome based region classification tasks

Dataset	Models (Order does NOT indicate anything)
Promoter_B_amyloliquefaciens	
Promoter_R_capsulatus	GROVER, HyenaDNA
Promoter_Arabidopsis_NonTATA	HyenaDNA
Promoter_Arabidopsis_TATA	HyenaDNA
human_vs_worm	Caduceus-Ph, GROVER, DNABERT-2
mouse_TFBS_1	NT-v2, DNABERT-2, GROVER
mouse_TFBS_2	Caduceus-Ph
mouse_TFBS_3	
mouse_TFBS_4	DNABERT-2
mouse_TFBS_5	NT-v2

For some datasets here (e.g. Promoter_GM12878 [lymphoblastoid cell line]), models achieve AUC around 96%-99%, indicating that zero-shot embeddings is already very powerful for such kind of tasks.

There is no uniformly best model

Multi-species genome based epigenetic medication identification

Element	Models (Order does NOT indicate anything)
A.thaliana_4mC	NT-v2, Caduceus-Ph
C.elegans_4mC	NT-v2, GROVER
D.melanogaster_4mC	NT-v2
E.coli_4mC	Caduceus-Ph, HyenaDNA
G.pickeringii_4mC	
G.subterraneus_4mC	
Yeast_H3	Caduceus-Ph, DNABERT-2
Yeast_H3K14ac	DNABERT-2, NT-v2
Yeast_H3K36me3	DNABERT-2, NT-v2
Yeast_H3K4me1	DNABERT-2
Yeast_H3K4me2	DNABERT-2
Yeast_H3K4me3	DNABERT-2
Yeast_H3K79me3	DNABERT-2
Yeast_H3K9ac	DNABERT-2, Caduceus-Ph
Yeast_H4	Caduceus-Ph, DNABERT-2
Yeast_H4ac	DNABERT-2

Human genome based epigenetic modification identification

Dataset	Models (Order does NOT indicate anything)
5mC	Caduceus-Ph, NT-v2
6mA	GROVER, NT-v2, Caduceus-Ph

When it comes to epigenetic identification, the AUCs got much lower (around 60%– 70%) for human genome and around 50%-60% for multi-species genome

Can DNA foundation model beat simple CNN

- We benchmarked the performance of these (zero-shot embedding + random forest) against a CNN trained from scratch as baseline.
- The CNN takes DNA sequence as input, one-hot encode it, and go through 3 convolutional layers.

Can DNA foundation model beat simple CNN

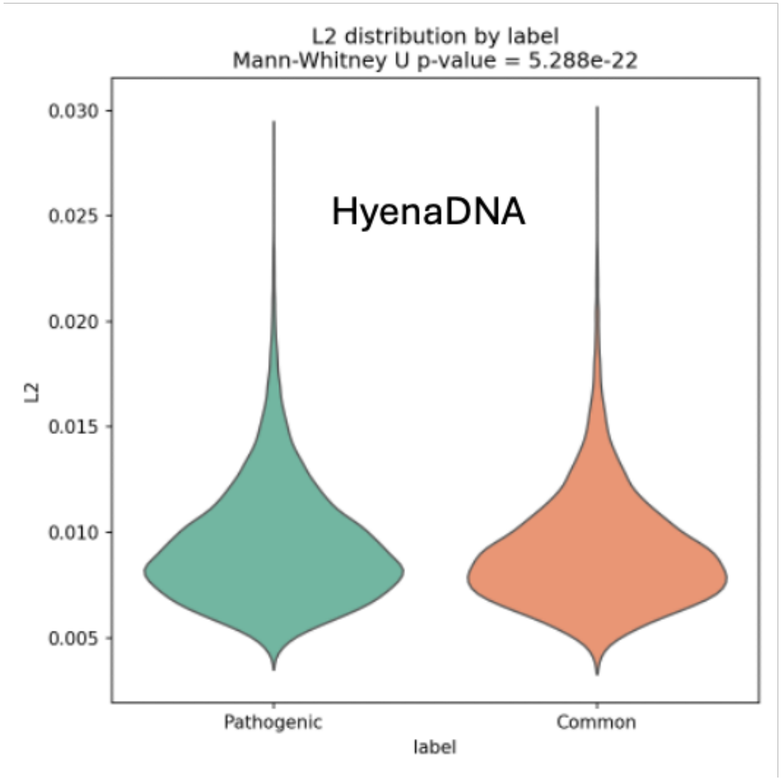
Dataset	Models
DNase_I	Caduceus-Ph
Human_TFBS_1	Caduceus-Ph, GROVER
Human_TFBS_2	Caduceus-Ph, GROVER
Human_TFBS_4	Caduceus-Ph
Promoter_GM12878	Caduceus-Ph, DNABERT-2, GROVER, NT-v2
Promoter_HUVEC	Caduceus-Ph, DNABERT-2, GROVER, NT-v2
Promoter_Hela-S3	Caduceus-Ph, DNABERT-2, GROVER, HyenaDNA, NT-v2
Yeast_H3	Caduceus-Ph, DNABERT-2
Yeast_H3K14ac	DNABERT-2, NT-v2
Yeast_H3K36me3	DNABERT-2, NT-v2
Yeast_H3K4me1	DNABERT-2
Yeast_H3K4me3	DNABERT-2
Yeast_H3K79me3	Caduceus-Ph, DNABERT-2, GROVER, NT-v2
Yeast_H3K9ac	Caduceus-Ph, DNABERT-2
Yeast_H4ac	DNABERT-2
coding	Caduceus-Ph
enhancer	DNABERT-2, GROVER, NT-v2
enhancer_cohn	Caduceus-Ph, DNABERT-2, GROVER, NT-v2
enhancer_ensembl	Caduceus-Ph

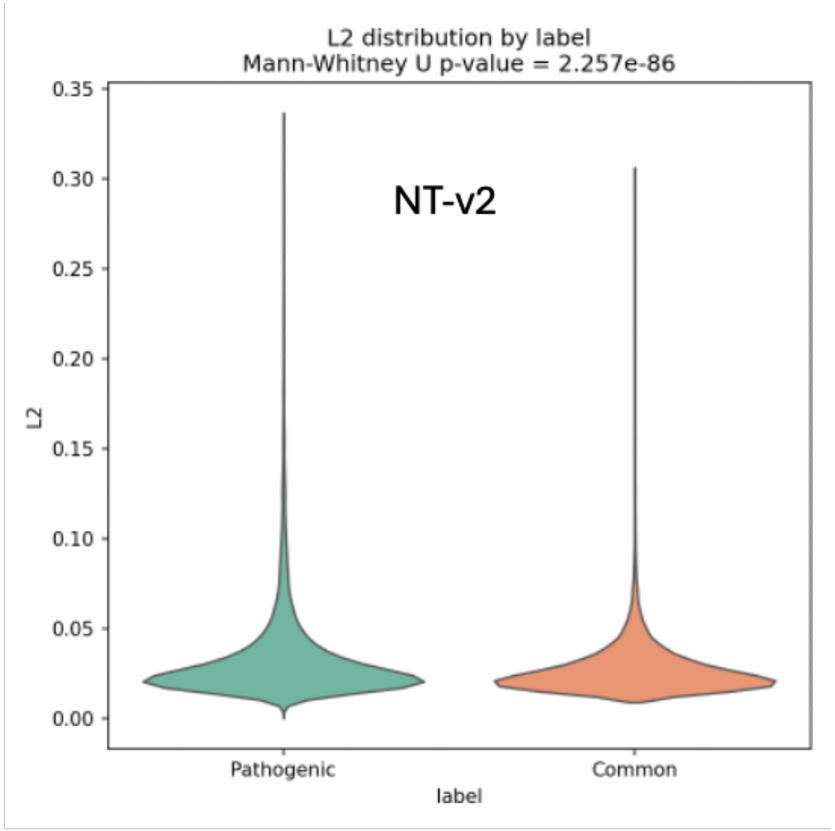
- Here are datasets where (zero-shot embedding + random forest) **outperforms** CNN.
- Besides the Yeast histone classification, we can see that all of them are human genome based region classification tasks.
- On the contrary, for almost all of the other datasets, the CNN beats almost all DNA foundation models.

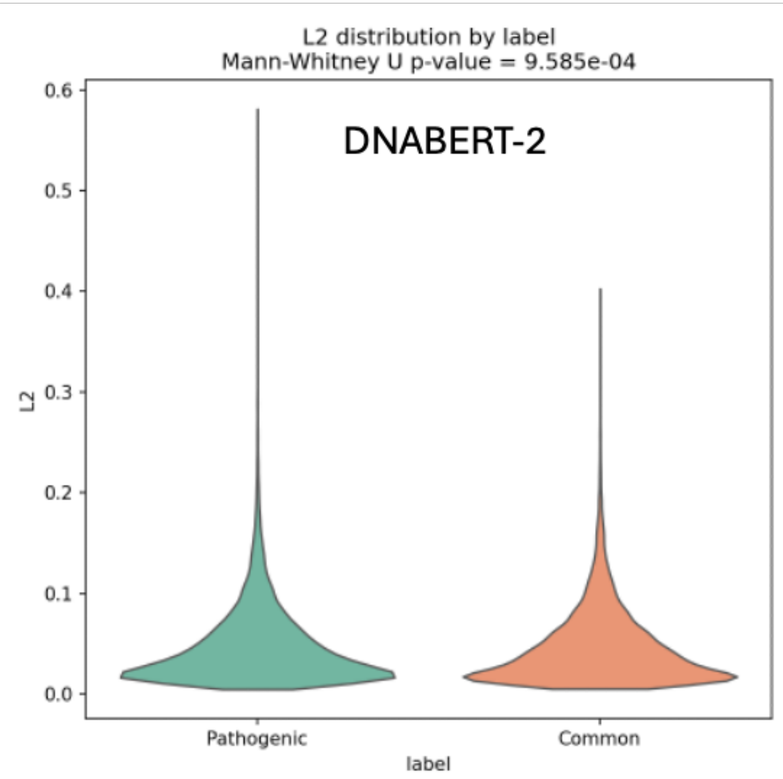
Benchmark on Genetic Variant Effect

- The embedding distance between a 6000 bps DNA sequence **with and without a pathogenic SNP** at its center, and
- The embedding distance between a 6000 bps DNA sequence **with and without a non-pathogenic (common) SNP** at its center
- The data are taken from InstaDeep's *genomic long range benchmark dataset (from HuggingFace)*. The sequences are generated from reference genome. We only consider chromosome 1, containing around **2000** pathogenic and **2000** common SNPs.
- We considered 4 different distance metrics: L1, L2, cosine similarity and dot product. We used the Wilcoxon rank-sum test to test the difference between the groups.
- Ideally, this two groups of embedding distances should be different.

Benchmark on Genetic Variant Effect



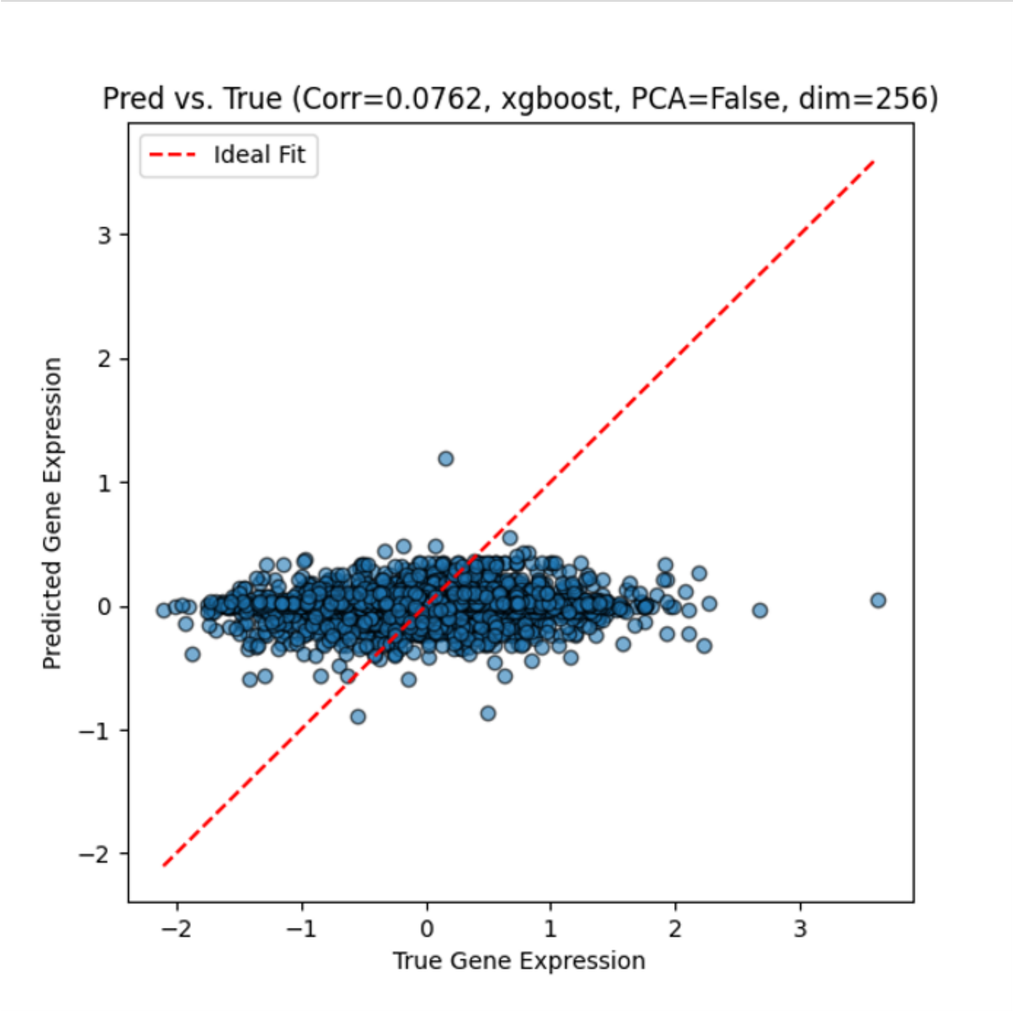




Can DNA foundation model predict gene expression

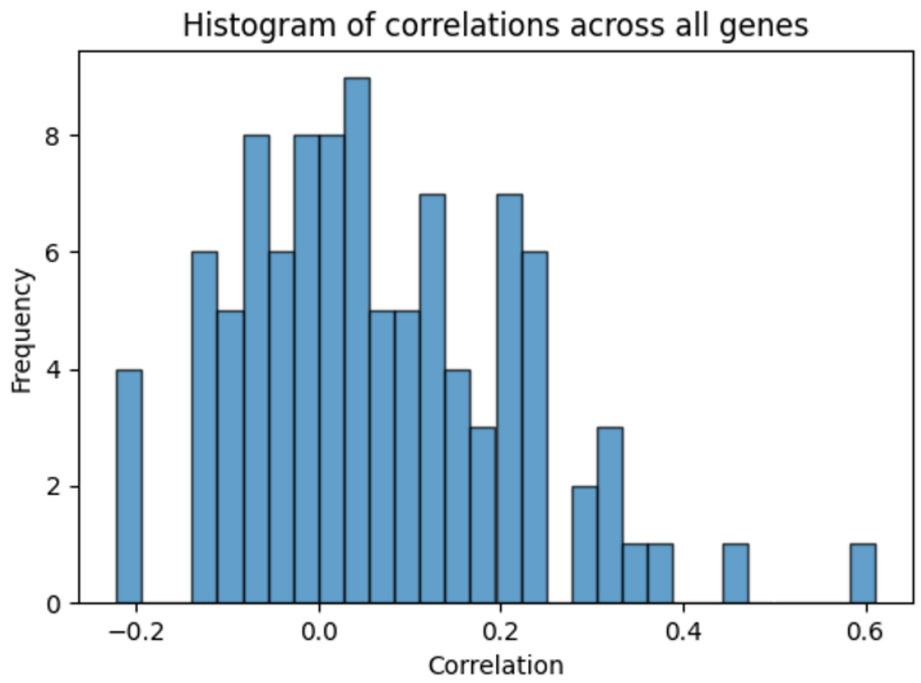
- We used GTEX v8 dataset to extract whole genome sequencing of **396** subjects, and their measured gene expression. We randomly chose **100** genes for preliminary exploration.
- This makes (DNA sequence, gene expression) for every subject, every gene. This is a **regression** task.
- The DNA sequence is defined as **6000 bps** around TSS for a gene. This is because NT-v2 and DNABERT-2 can handle sequence up to this length. GROVER can handle 512 tokens (around 2500 bps), which is too short for such task, so we excluded it.

Can DNA foundation model predict gene expression



- We train all the genes together (either fine tune or zero shot embedding).
- This is the scatterplot (true vs prediction in test set) of using HyenaDNA. The scatterplots are very similar for all other DNA foundation models & experiments.
- The overall performance is not very excited partly because the heritability of gene expression is often low.

Can DNA foundation model predict gene expression



When train each gene separately, the results are slightly better, but still not very exciting.

Discussion

- Many more models are publicly available; we want to particularly mention one: Evo2 (which may be state-of-the-art due to much larger training samples and parameters): a smaller version at 7B parameters trained on 2.4 trillion tokens and a full version at 40B parameters trained on 9.3 trillion tokens; Hyena based structure
- Downstream task requires specific considerations. For example, in gene expression prediction task, carefully designed objective function for fine tune or further full train may be a promising way to go (Performer vs Enformer)
- Open science and easy to use are key for wide adoption by the community: HuggingFace is really good place to deposit data and model (the original version of the data in this work is in OSF.io, we will deposit all resources and curated datasets into HuggingFace)

Acknowledgements

- Haonan Feng did most of the work and conducted all the experiments
- The work is co-led by Peng Wei
- Thank Bingxin for invitation and involvement in this work!
- We thank NIH and CPRIT for support

Thank you!

Chong Wu

Email: cwu18@mdanderson.org

Website: https://wuchong.org