LMM for Gene Repression Correlation

Data

- Rabani_Expression_A+

id	1h	2h	3h	4h	5h	6h	7h	8h	10h
S0_M_T1	1.1502	1.1256	1.4005	0.2332	0.73195	-0.47038	-0.57411	-0.25983	-0.76564

- Mean RBP Data

Motif	Mean_RBP
TGTGGAT	7.17826412

Theory

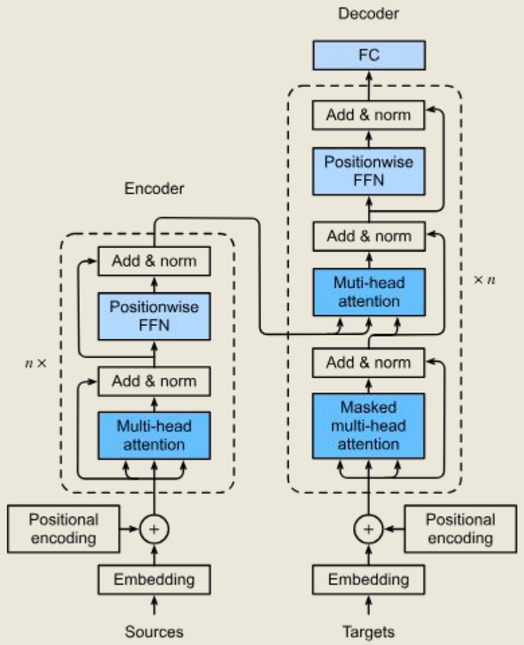
- Given the ‘Motif Composition’ of a sequence, we can predict it’s Repression Rate
- Motif Composition: Motifs that make up a sequence; motifs with higher Mean RBP have higher significance
- We can measure correlation using LLM

GGGTCCCCCTGACAG

Methodology

GGGTCCCCCCTGACAG ➡ TOKENIZER

[GGG], [TCCCCCCT], [GACAG] ➡



➡ Prediction (Scalar)

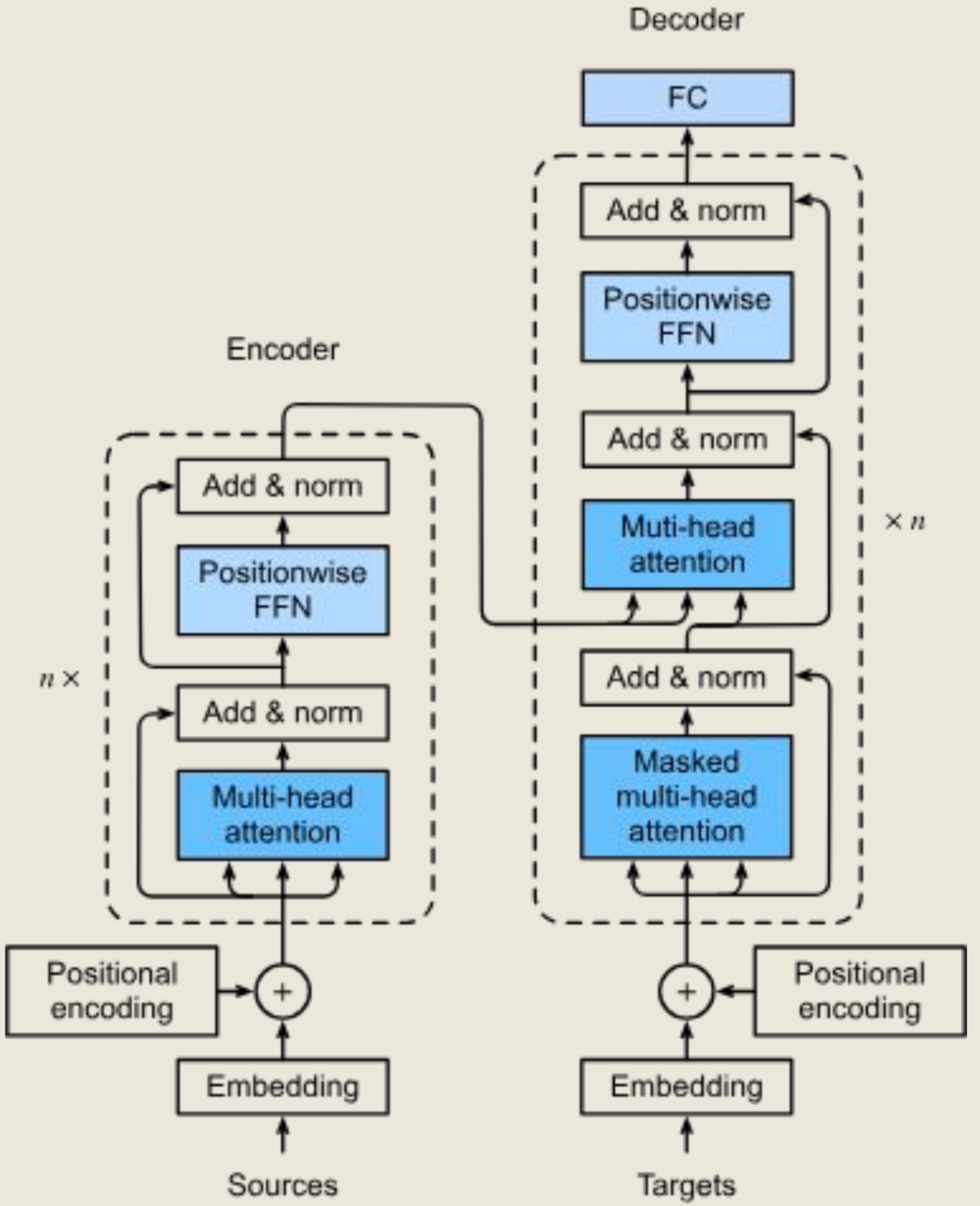
Large Language Model

Originally made for translating languages.

Takes tokens as numbers and performs matrix operations to achieve output

Has no concept of ‘Order’ or ‘Characters’, only tokens

Fine Tuning: Utilizing pretrained model for quick learning



Experimentation with Transformers

DNA-MORPH

- BERT-Model from scratch
- Transformer from scratch
  - Ranked motifs by Mean RBP
  - Found ‘words’ by Motif Ranking according to Morphological Segmentation

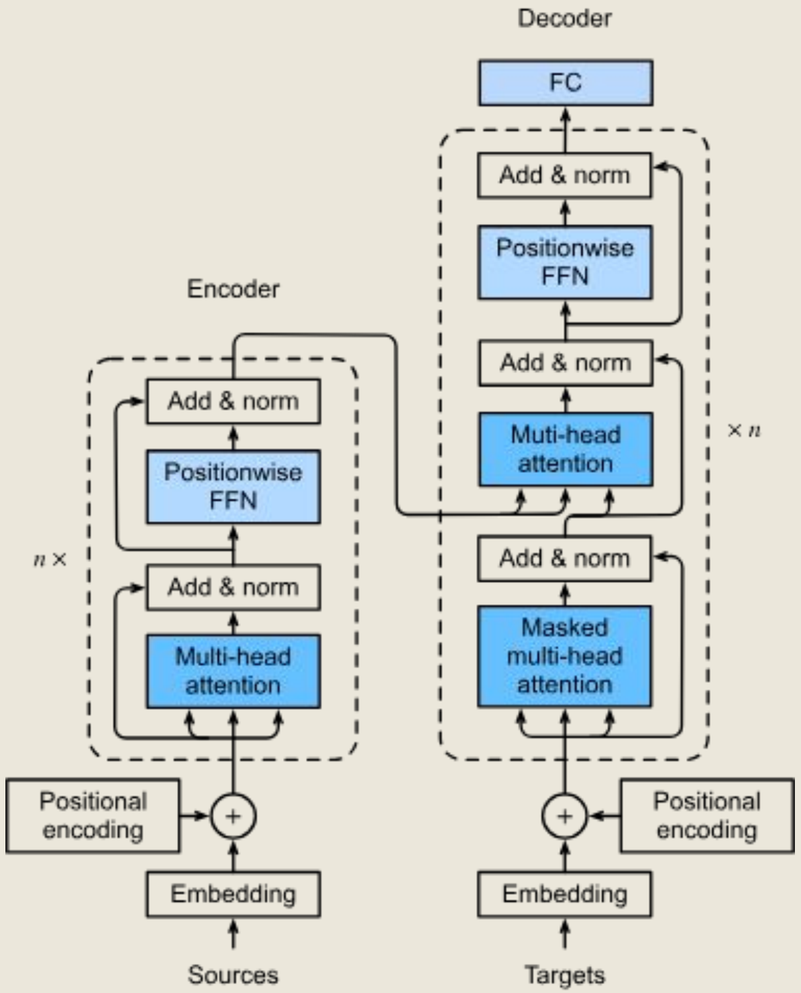
MORPH_Dmitri	MORPH_Owen
• Trained on Dmitri Data	• Trained on Owen Data

Experimentation with Transformers

Do LLMs Work?

DNABERT2:

- r2 = .335



Experimentation with Transformers

DNAMORPH has 2 functions:

- Segment Sequences into Motifs:

TGTCCCC: 12.12

GTCCCCG : 1.01

TCCCCGG: 4.43

TGTCCCCGGGTCTT→[TGTCCCC, TCCCCGG,...]

- ‘Tokenize’ Motifs :

[TGTCCCC, TCCCCGG,...] → [123233, 13332, ...]

Experimentation with Transformers

Seq = TGTCCCCGGGTCTTCCAACGGACTGG...

Sliding Window(Seq) = GGGTCTT CACTGGG ACTGGGG...

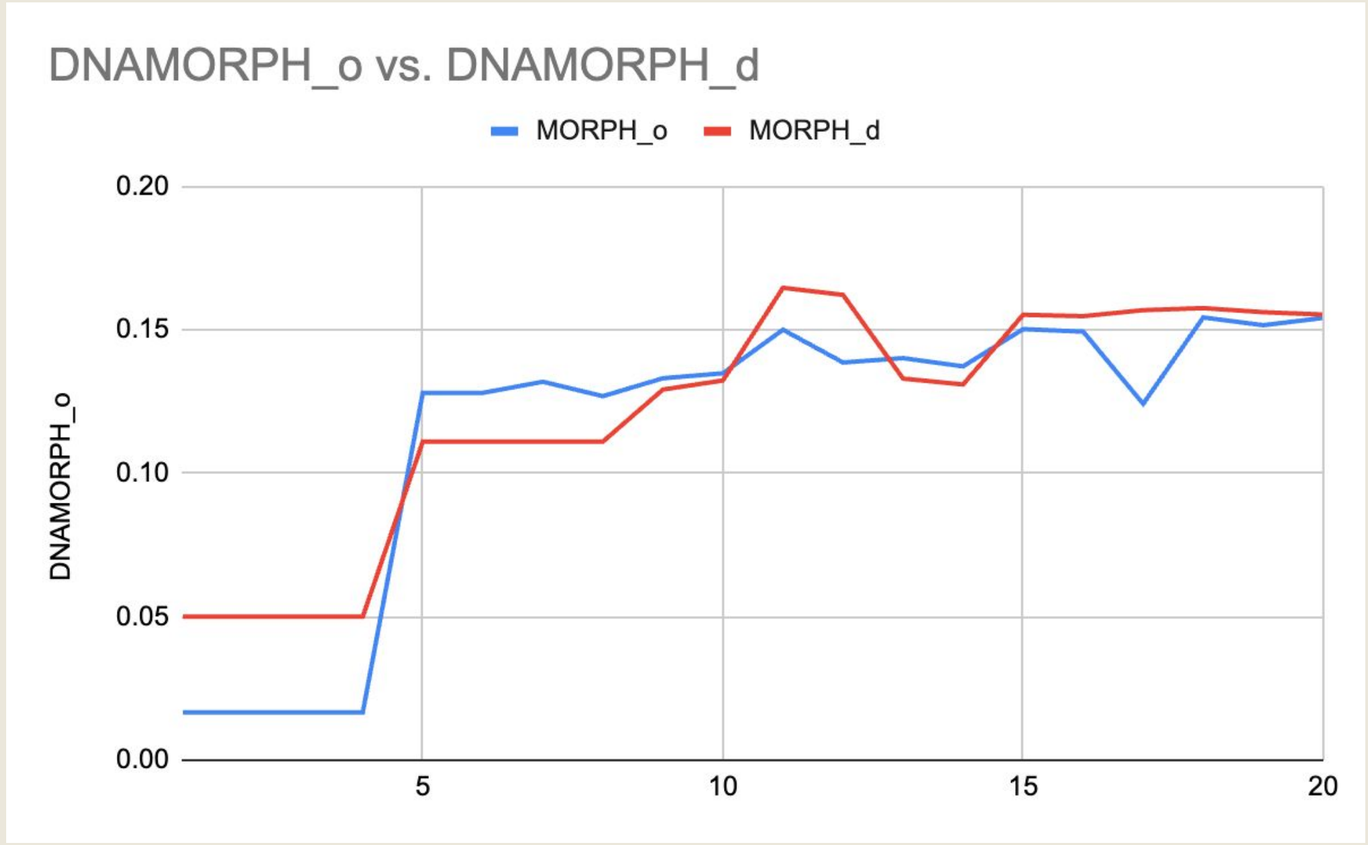
- MORPH_O:

  ['gggtctt', 'cactggg', 'actgggg']

- MOPRH_D
  - ['gggtctt', 'cactgg', '##g', 'actgg', '##gg']

Experimentation

- MORPH_O
  - R^2 = .178
- MORPH_D
  - R^2 = 0.174



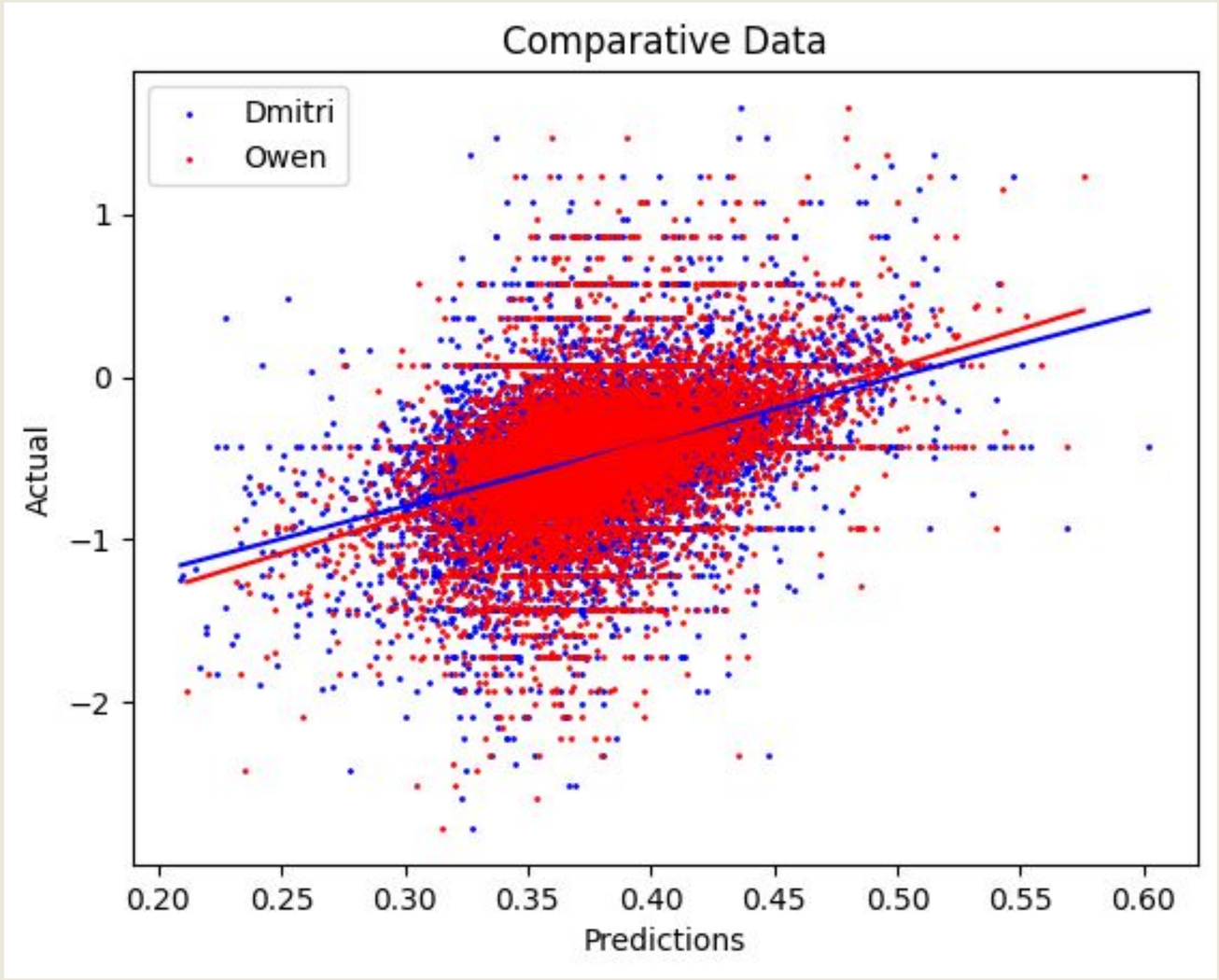
R2: Proportion of variance in the dependent variable that can be explained by the independent variable

Measuring Correlation with Models

	Model_d	Model_o
Mean Squared Error	.144	.141
Mean absolute error	.265	.261
Root Mean Squared Error	.379	.375

Measuring Correlation with Models

Compare this with canonical data to see which data gives stronger predictions, thus has more correlation



Measuring Correlation with Models

	Model_d	Model_o	Model_DNA
Mean Squared Error	.144	.141	0.092
Mean absolute error	.265	.261	.216
Root Mean Squared Error	.379	.375	.304
R^2	.174	.178	.457

Measuring Correlation with Models

As compared to DNABERT2