Applying machine learning to biological promoters

Nathan Hu

What is the research problem

Project overview

- Largely based off two papers applying machine learning to promoter prediction

ARTICLE

https://doi.org/10.1038/s41467-020-15977-4 OPEN

Model-driven generation of artificial yeast promoters

Benjamin J. Kotopka[1] & Christina D. Smolke[1,2]✉

Promoters play a central role in controlling gene regulation; however, a small set of promoters is used for most genetic construct design in the yeast *Saccharomyces cerevisiae*. Generating and utilizing models that accurately predict protein expression from promoter sequences would enable rapid generation of useful promoters and facilitate synthetic biology efforts in this model organism. We measure the gene expression activity of over 675,000 sequences in a constitutive promoter library and over 327,000 sequences in an inducible promoter library. Training an ensemble of convolutional neural networks jointly on the two data sets enables very high ($R^2 > 0.79$) predictive accuracies on multiple sequence-activity prediction tasks.

Genome analysis

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

Yanrong Ji[1,†], Zhihan Zhou[2,†], Han Liu[2,*] and Ramana V. Davuluri[3,*]

[1]Division of Health and Biomedical Informatics, Department of Preventive Medicine, Northwestern University Medicine, Chicago, IL 60611, USA, [2]Department of Computer Science, Northwestern University, Evanston, I
[3]Department of Biomedical Informatics, Stony Brook University, Stony Brook, NY 11794, USA

*To whom correspondence should be addressed.
[†]The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.
Associate Editor: Dr. Janet Kelso

Received on September 10, 2020; revised on December 31, 2020; editorial decision on January 25, 2021; accepted on February 1, 2021

Abstract

Motivation: Deciphering the language of non-coding DNA is one of the fundamental problems in g
Gene regulatory code is highly complex due to the existence of polysemy and distant semantic re
previous informatics methods often fail to capture especially in data-scarce scenarios.

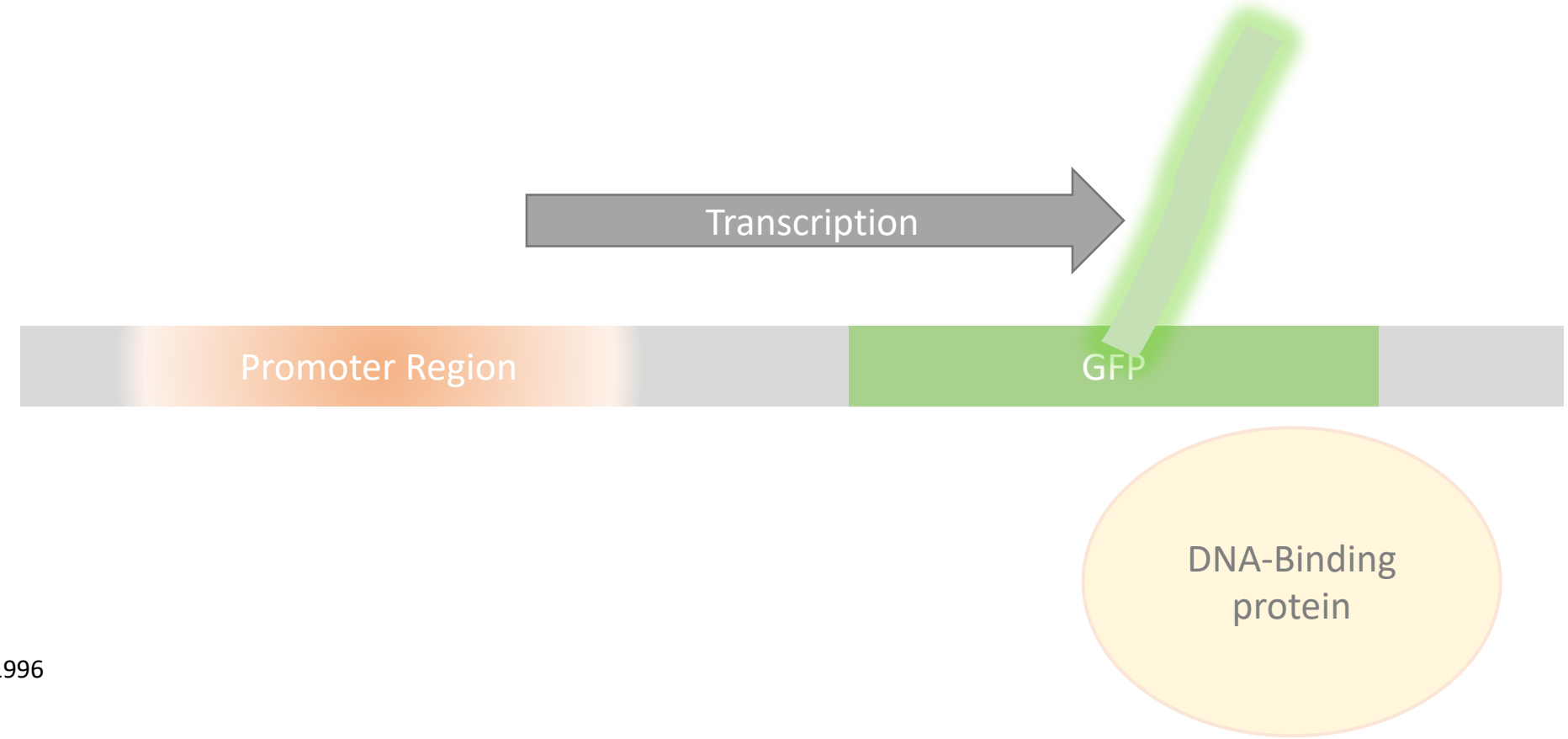
What is a promoter

- Promoters are the key non-coding regions of the genome[2]
- Found near the transcription start site of a gene[2]
- Important for initiation of transcription and for regulating gene expression[1] by serving as recognition sites for necessary proteins[2]

[1]Ayoubi & Van De Yen 1996

[2]Oubounyt et al. 2020

Green Fluorescence Protein (an example)



[1]Ayoubi & Van De Yen 1996

Motivation

- Precise gene expression control is required in[3]:
  - Engineered metabolic pathways
  - Gene circuits

- Finding promoters or creating artificial promoters with useful properties may aid gene construct design[3]

- Studying DNA with machine learning may aid understanding of promoters and DNA regulation in general

[3]Kotopka & Smolke 2020

Question:

How can we use machine learning to make predictions on promoters?

Research question

How can we use machine learning to make predictions on promoters?

Regression

How strong is this promoter?

(number value)

Example: Creating artificial promoters

ATGTAAGTGACASTARTGACAGGACATAGACATTACATAT

Unknown New Sequence

Model says it's a strong promoter

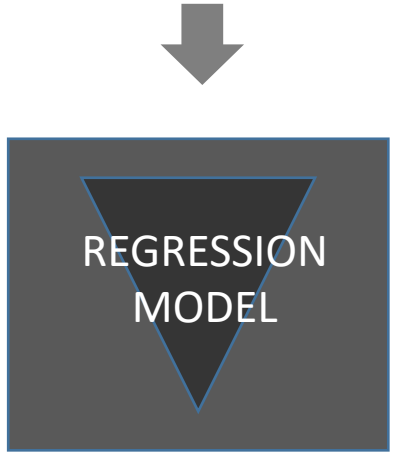
Verify model prediction *in vivo*

New promoter sequence found!

Goal

AGGACATAGACATTACATAT

Input DNA sequence



Blackbox model

0.67

Model output
(promoter strength)

DNA vector representation

DNA is a sequence of 4 base pairs

How do we represent a sequence such as ATTGCT in a computer readable way?

One-hot encoding

Each base represented by a vector:

$$A = \begin{pmatrix}1\\0\\0\\0\end{pmatrix}, \quad G = \begin{pmatrix}0\\1\\0\\0\end{pmatrix}, \quad C = \begin{pmatrix}0\\0\\1\\0\end{pmatrix}, \quad T = \begin{pmatrix}0\\0\\0\\1\end{pmatrix}$$

A sequence is represented by a 2D matrix:

ATTGCT becomes

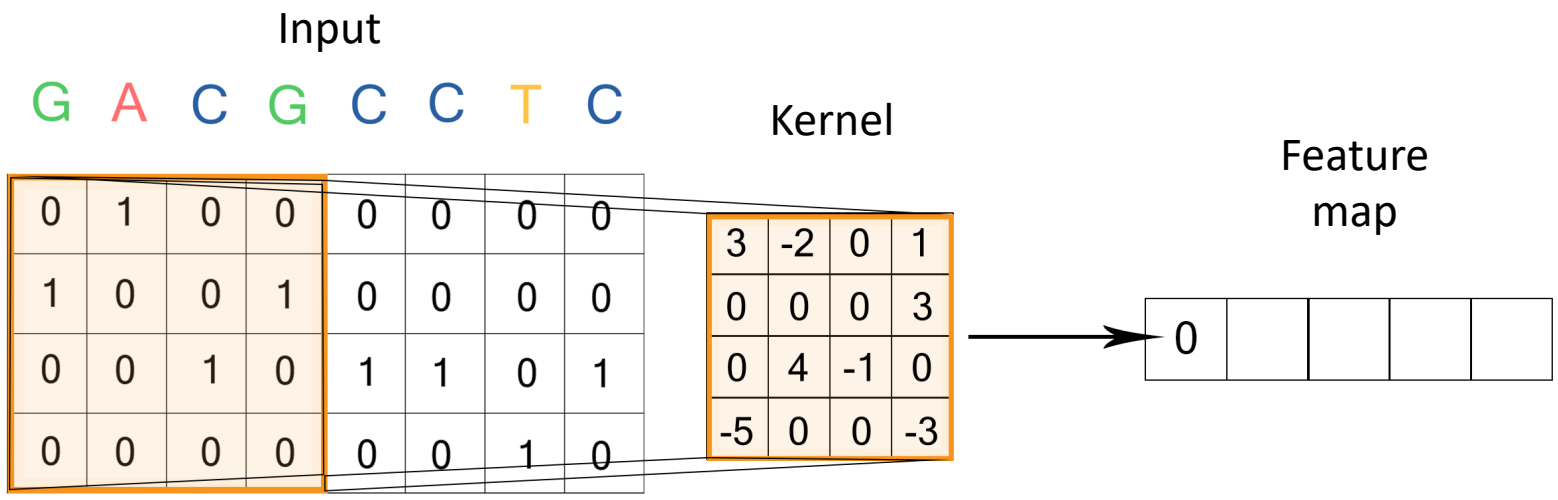
A T T G C T

$$\begin{bmatrix}100000\\000100\\000010\\011001\end{bmatrix}$$

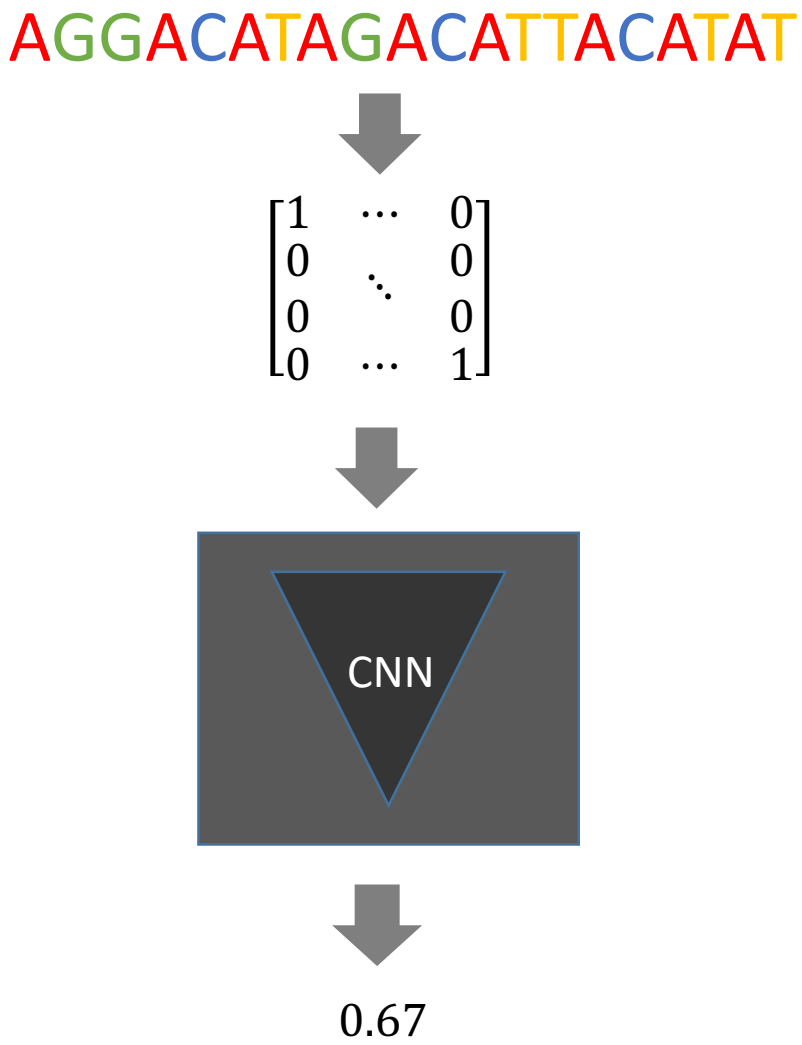
What architectures can we use?

Convolutional neural network (CNN)

- Usually applied to images and video
- Based off the concept of a convolution
- Focuses on the input 1 section at a time



Pipeline 1



Input DNA sequence

One-hot encoding

Blackbox model

Model output
(promoter strength)

How else can we represent a sequence

We have one-hot encoding

Another approach is to learn a representation for DNA

[5]

DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome

[5]Ji et al. 2021

DNABERT[5]

Based off BERT[7], a state-of-the-art natural language model

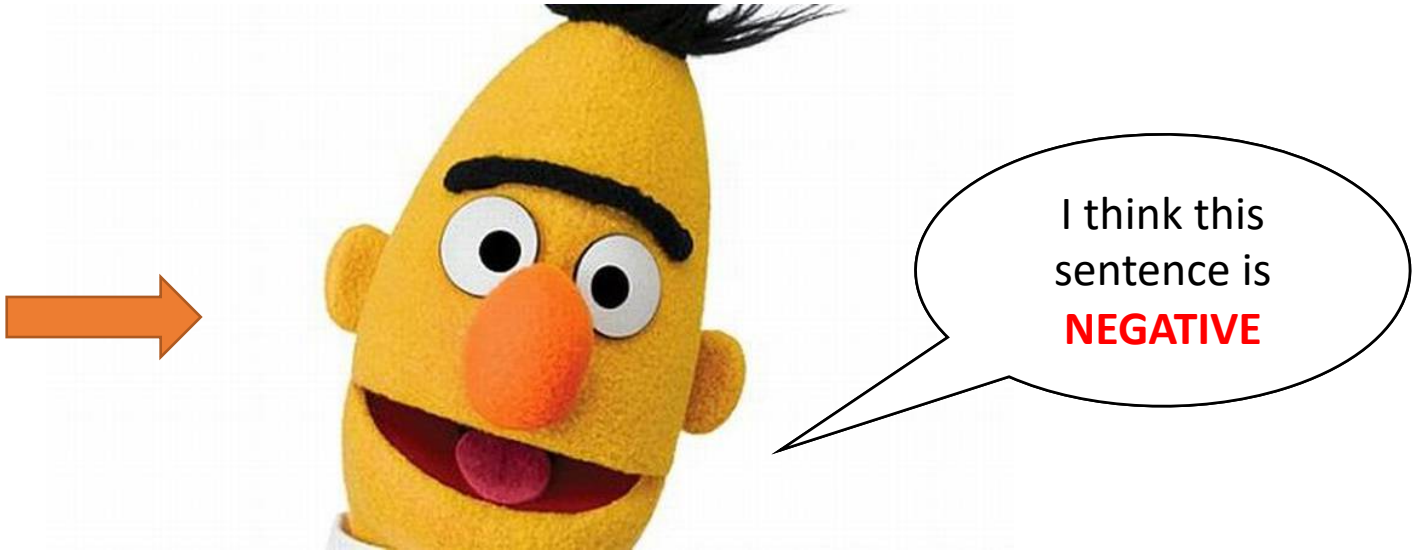
Able to take in sentences as input and make different predictions

[7]Devlin et al. 2018

BERT[7] Language Example

Sentiment analysis[6]

"I hate you"



BERT

[6]Hugging Face 2020
[7]Devlin et al. 2018

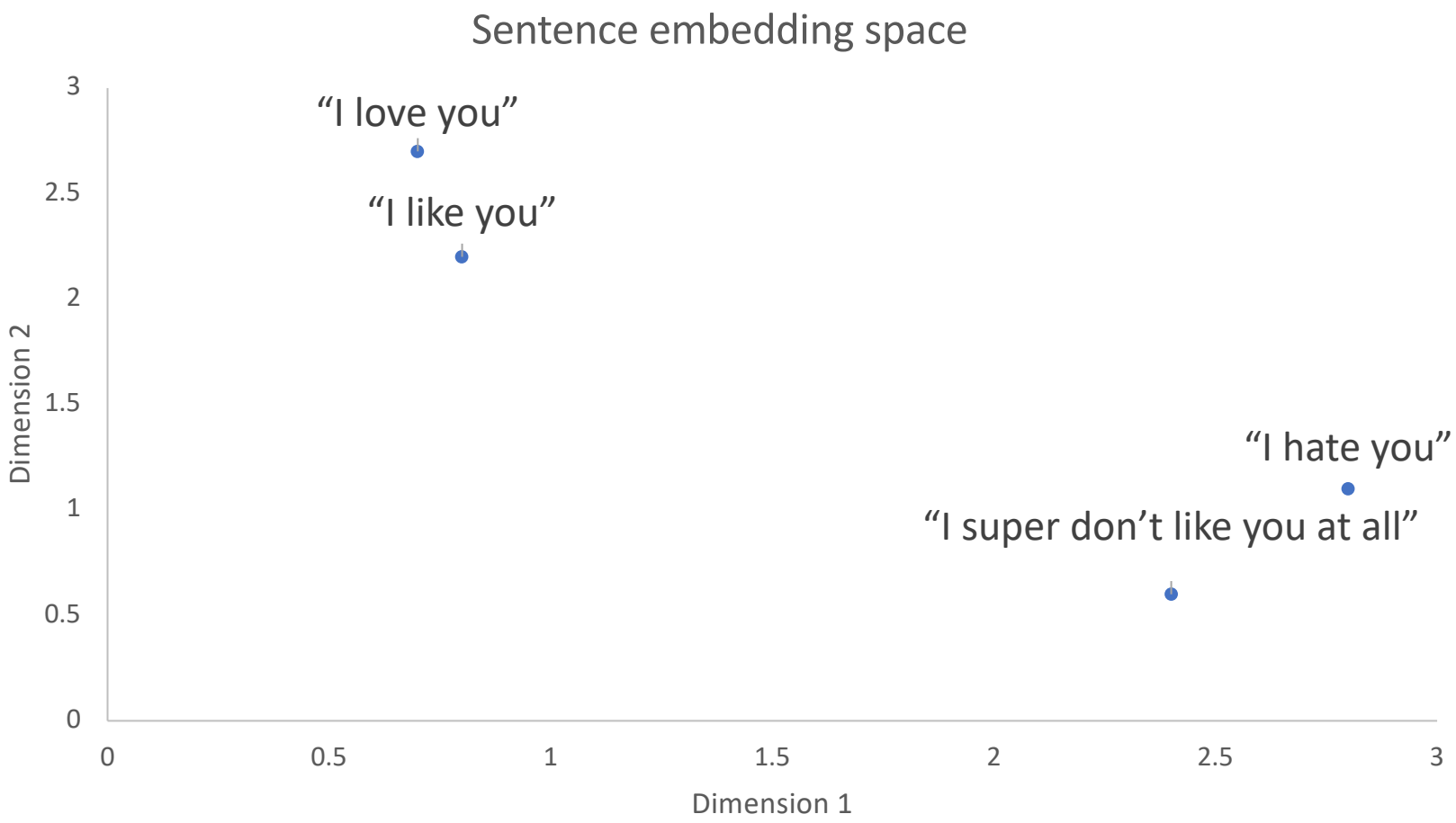
Sentence representations

BERT can ‘embed’ or represent each sentence as a vector

“I hate you” → $[1.3221, -11.823, 0.332 \cdots 2.398]$

“I love you” → $[-3.928, 0.991, 6.122 \cdots 7.082]$

Sentence representations



Attention mechanism

BERT can separate different sentences by understanding context

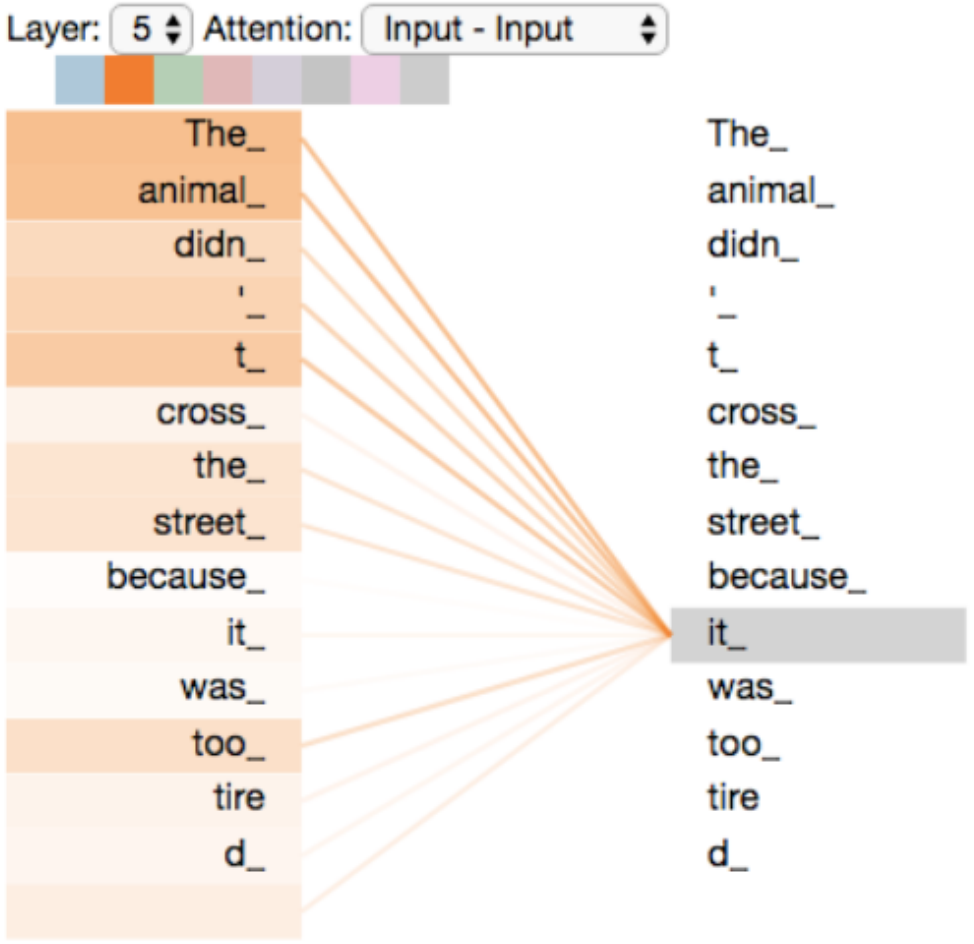
Understands context by using the “attention mechanism” on each word

“I love you”

“I”

“love”

“you”



[4]Alammar 2018

Link to DNA

A sequence of DNA is analogous to a sentence

GACAGGACATAGACATTACAT is a ‘sentence’

DNABERT

DNABERT[5] embeddings are first trained using the human genome for 25 days on 8 GPUS

Provides it with a general and transferrable understanding of DNA[5]

So can be transferred to different downstream tasks and organisms

Yeast promoter strength prediction

[5]Ji et al. 2021

One-hot vs DNABERT

AGGACATAGACATTACATAT

One-hot encoding

$$\begin{bmatrix} 1 & \cdots & 0 \\ 0 & \ddots & 0 \\ 0 & & 0 \\ 0 & \cdots & 1 \end{bmatrix}$$

DNABERT learned embedding

$$[1.3221, -11.823, 0.332 \cdots 2.398]$$

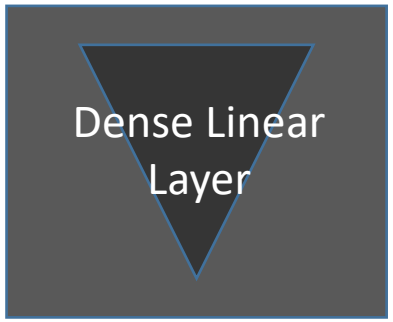
Pipeline 2

AGGACATAGACATTACATAT

Input DNA sequence

$[1.3221, -11.823, 0.332 \cdots 2.398]$

DNABERT embedding

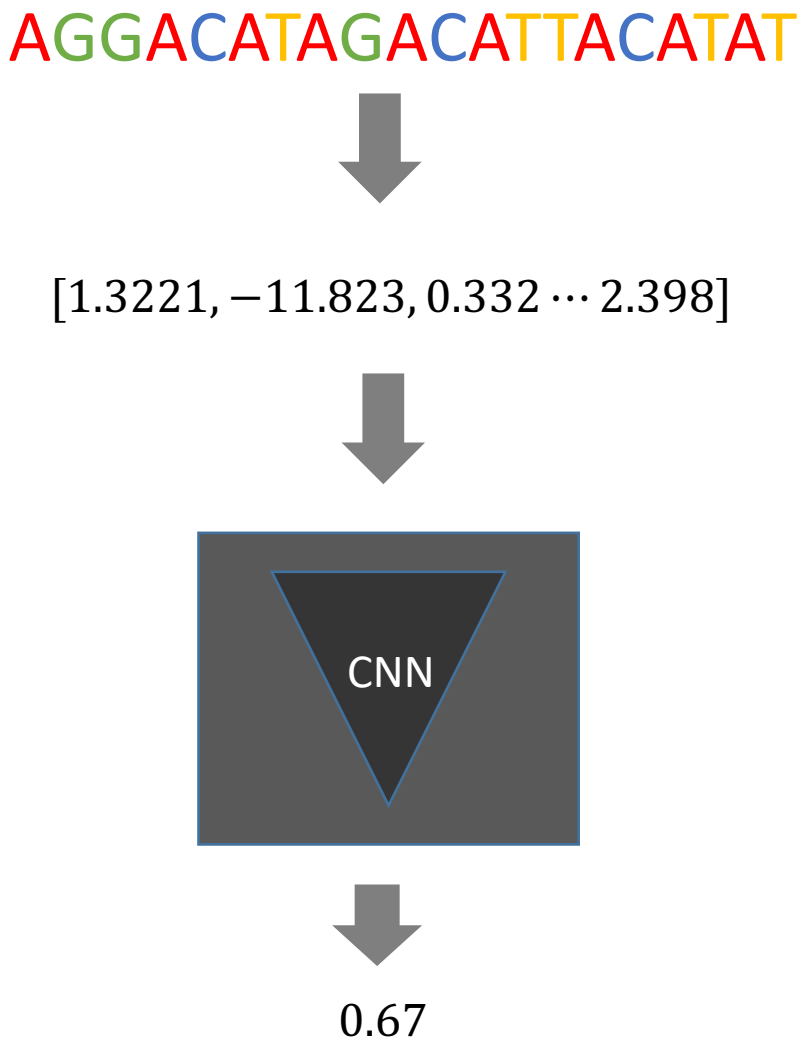


Blackbox model

0.67

Model output
(promoter strength)

Pipeline 3



Input DNA sequence

DNABERT embedding

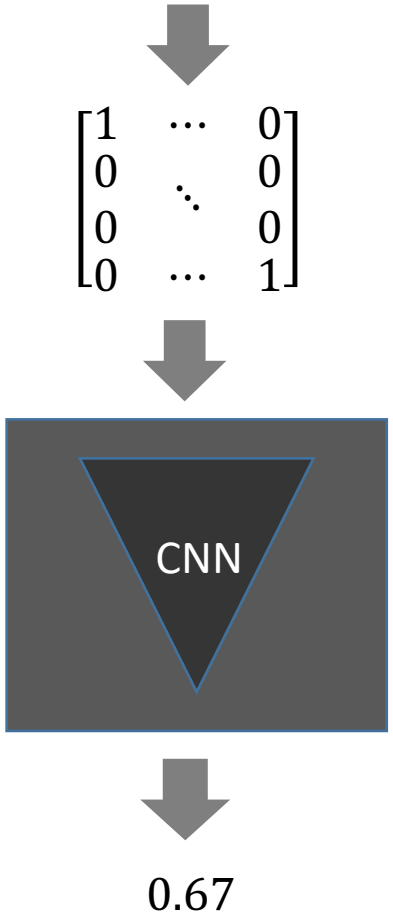
Blackbox model

Model output
(promoter strength)

Summary of architectures

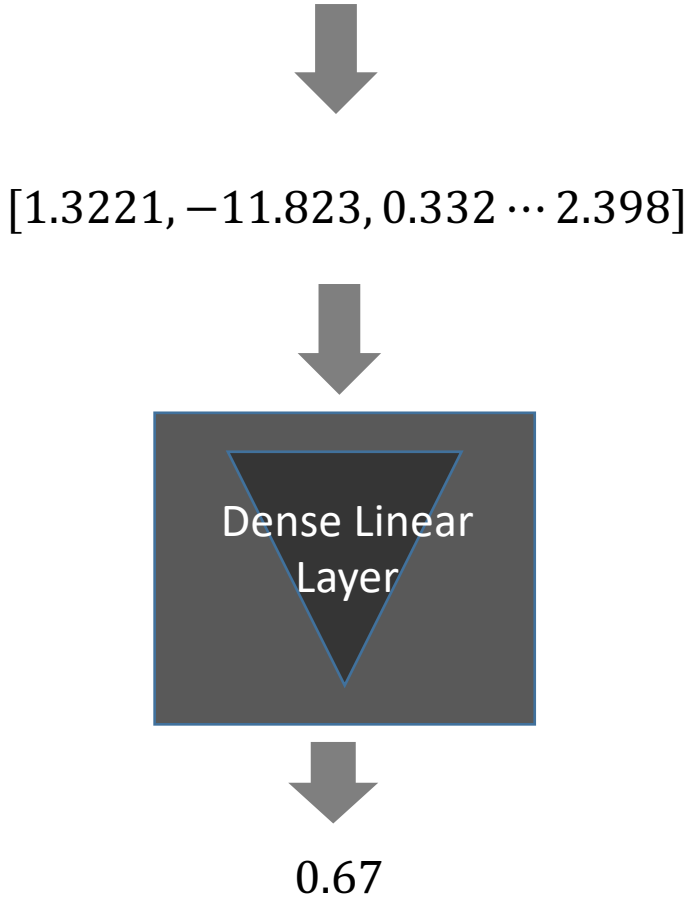
1. One-hot + CNN

AGGACATAGACATTACATAT



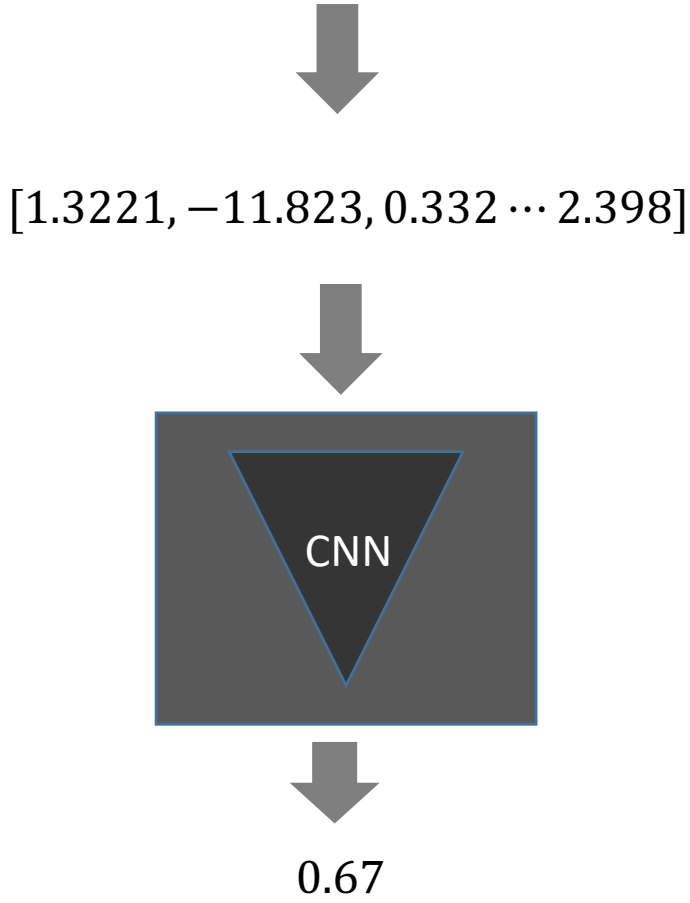
2. DNABERT + Dense

AGGACATAGACATTACATAT



3. DNABERT + CNN

AGGACATAGACATTACATAT

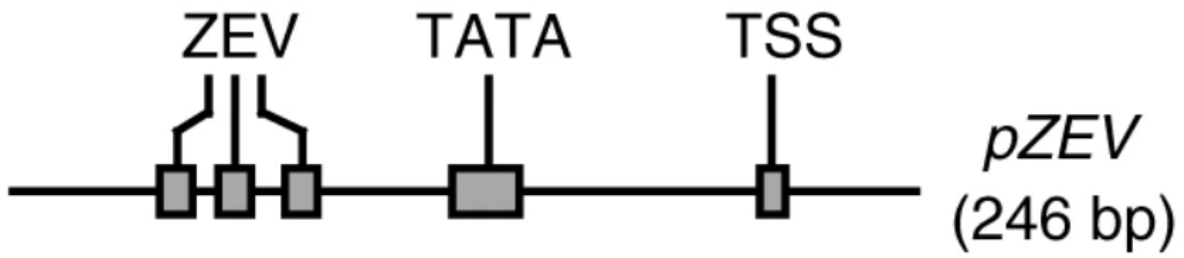


Methods & Results

The dataset[3]

327,000 yeast promoters containing a binding site for artificial transcription factor ZEV

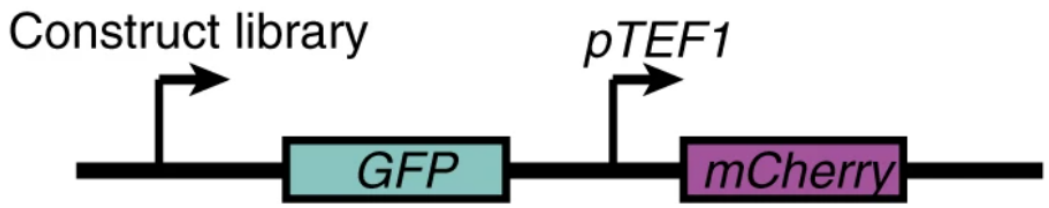
Each sequence is 246 base pairs and has a label for measured promoter strength



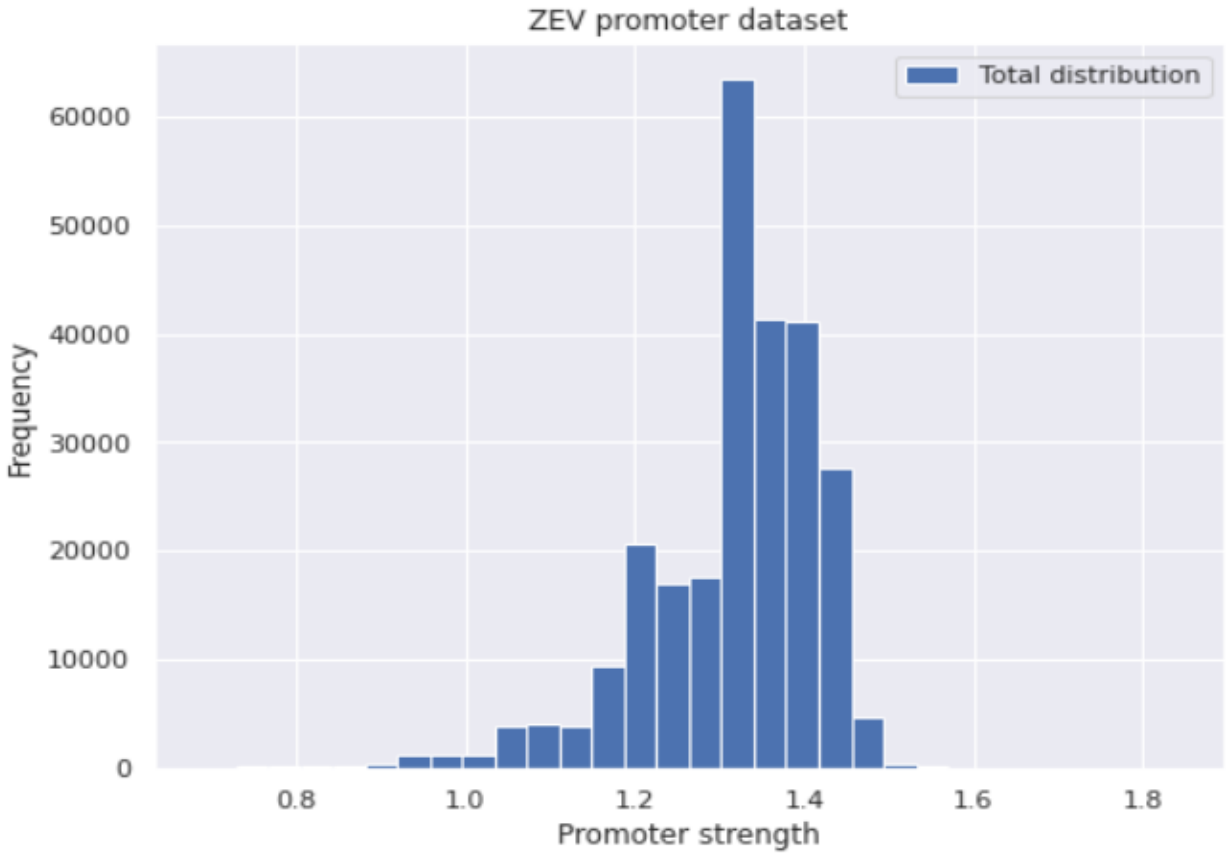
[3]Kotopka & Smolke 2020

The dataset

How was promoter strength measured?



[3]Kotopka & Smolke 2020



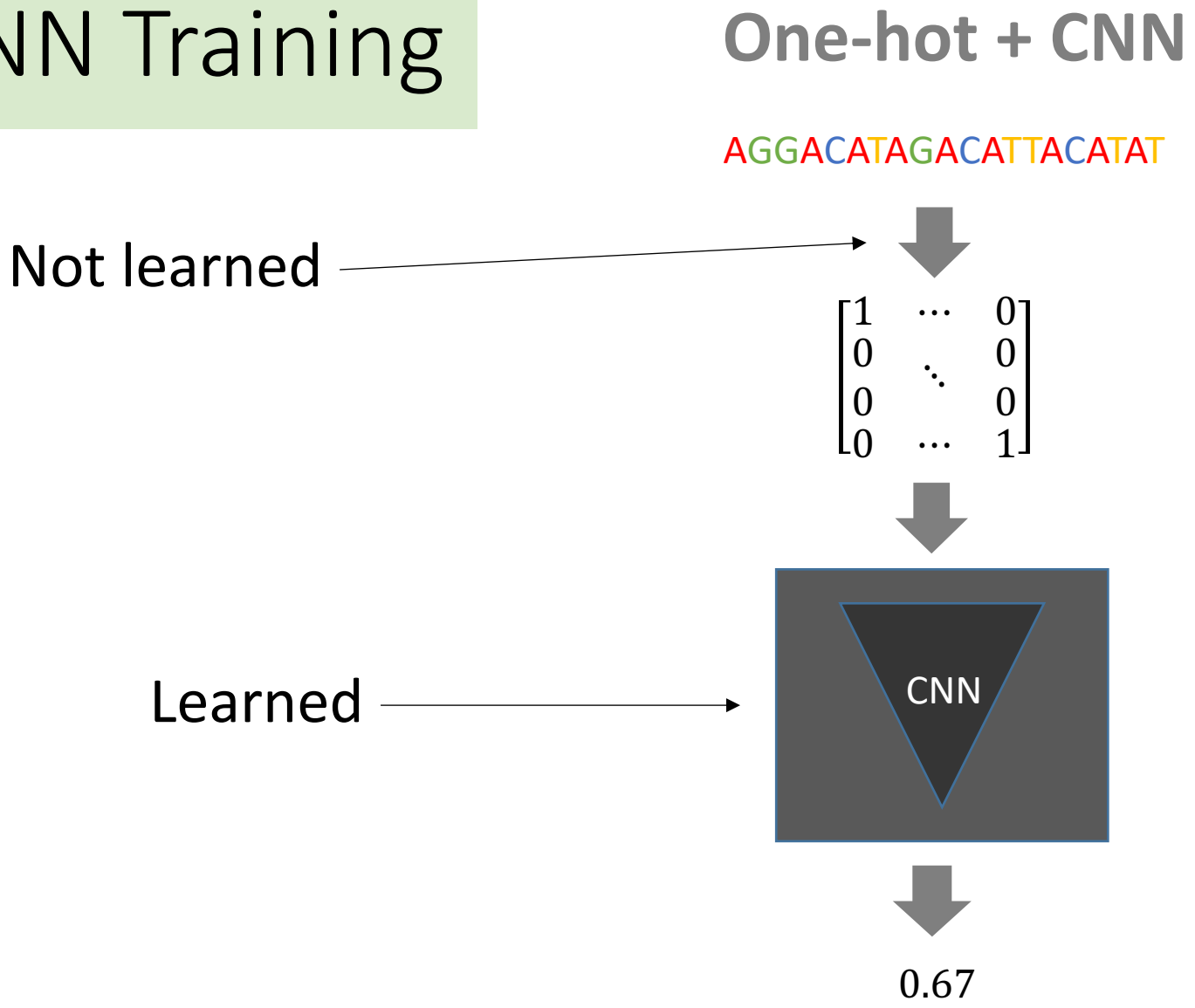
Training

Each model needs to be trained to make accurate predictions

We split the 327,000 promoters into

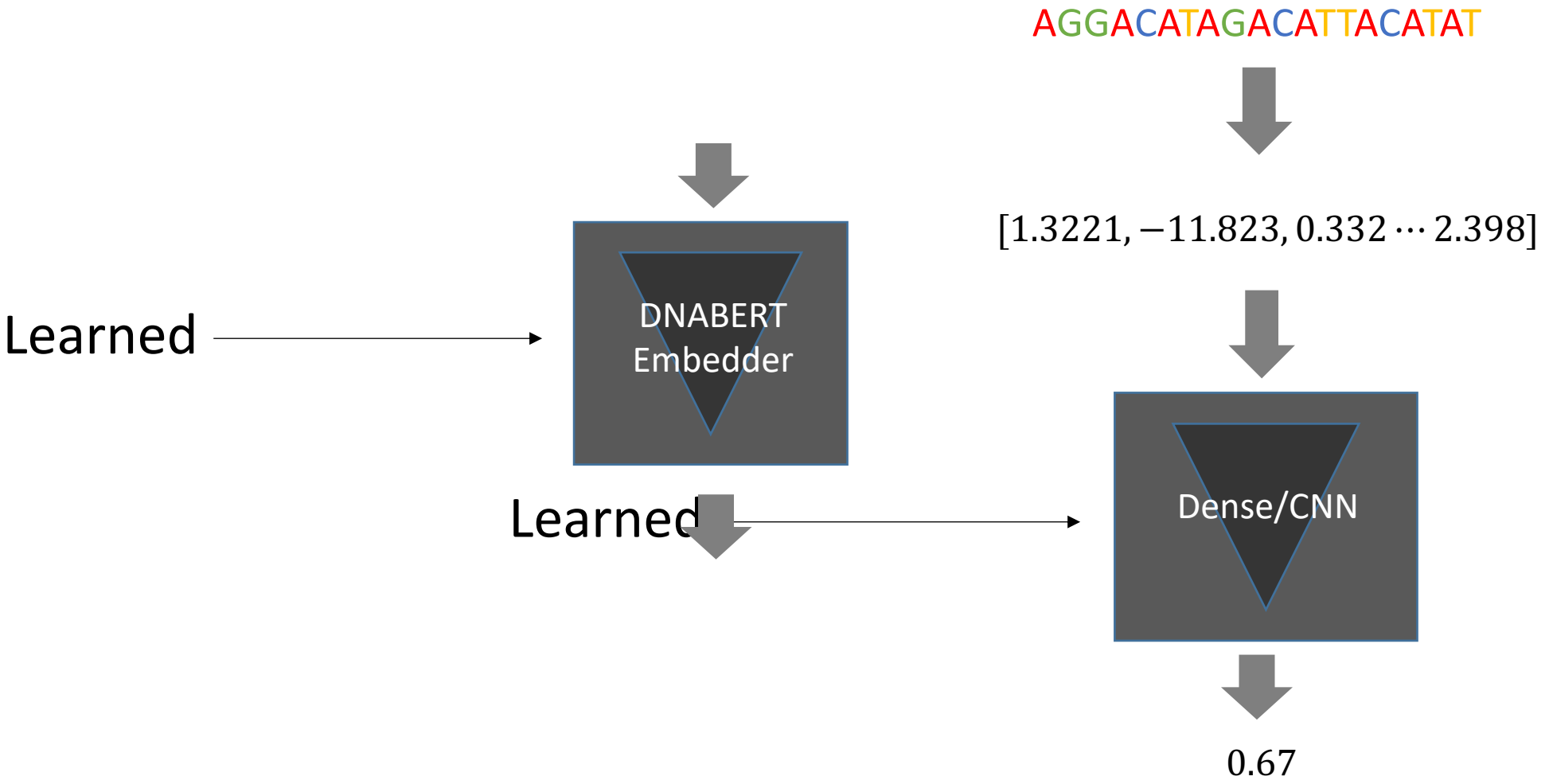
- 261,584 **training** examples
- 32,708 **validation** examples
- 32,708 **test** examples

One-hot+CNN Training

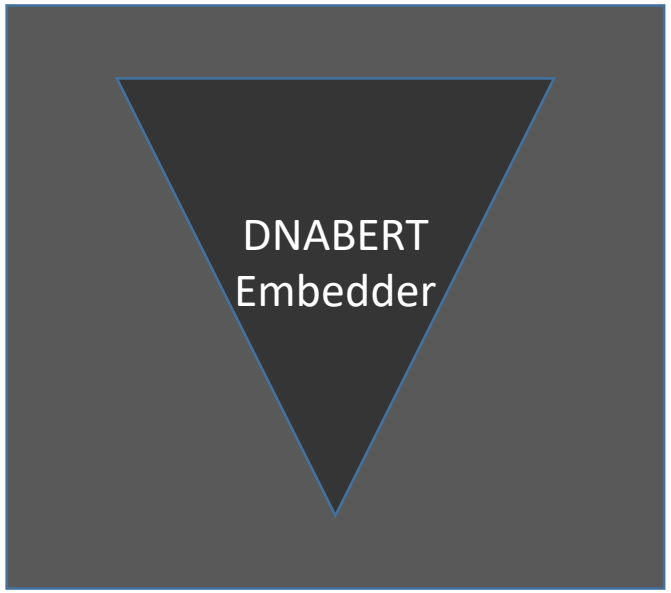


DNABERT Training

DNABERT + Dense/CNN



DNABERT Training



Very big!!

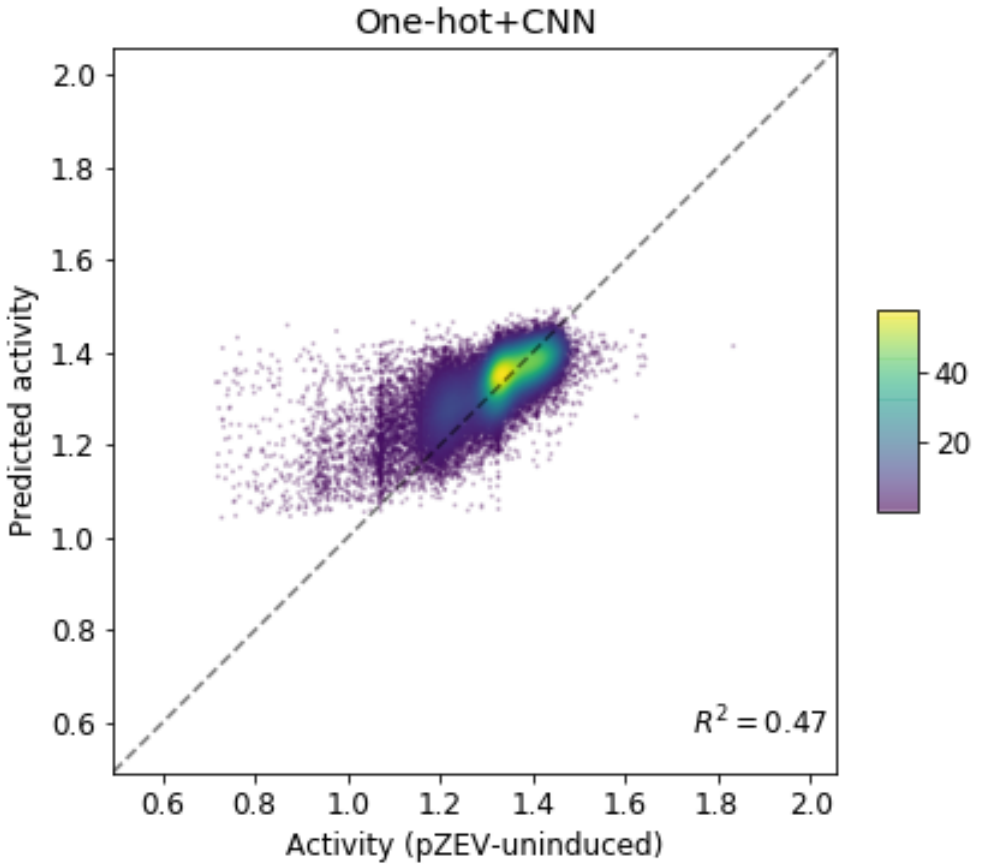
DNABERT has lots of layers

We download a pre-trained version that is trained using the entire **human** genome

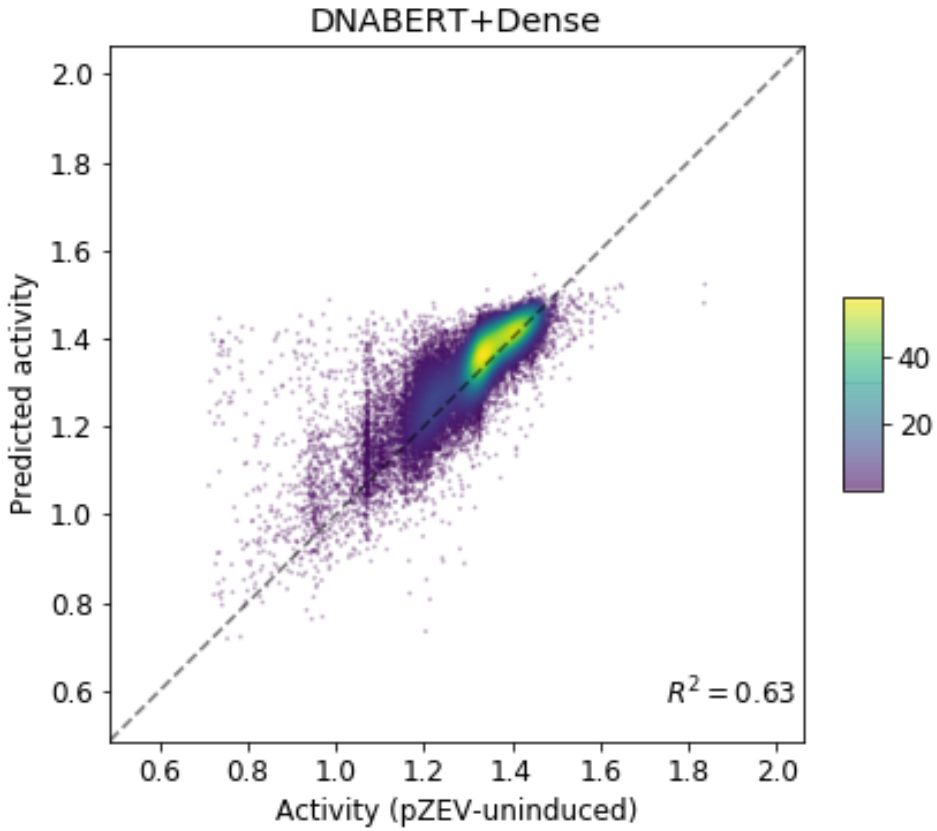
We then make fine adjustments using the **yeast** promoter dataset

1. One-hot + CNN

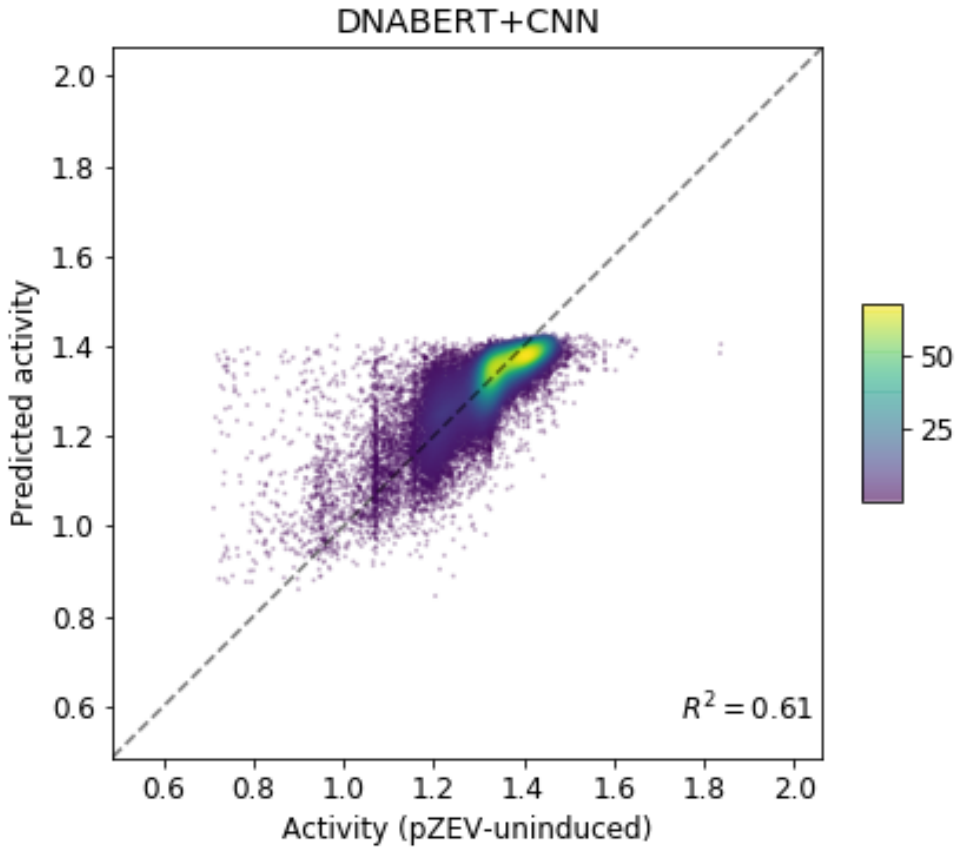
1. One-hot + CNN


One-hot+CNN
Predicted activity
Activity (pZEV-uninduced)
2.0
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
$R^2 = 0.47$
40
20

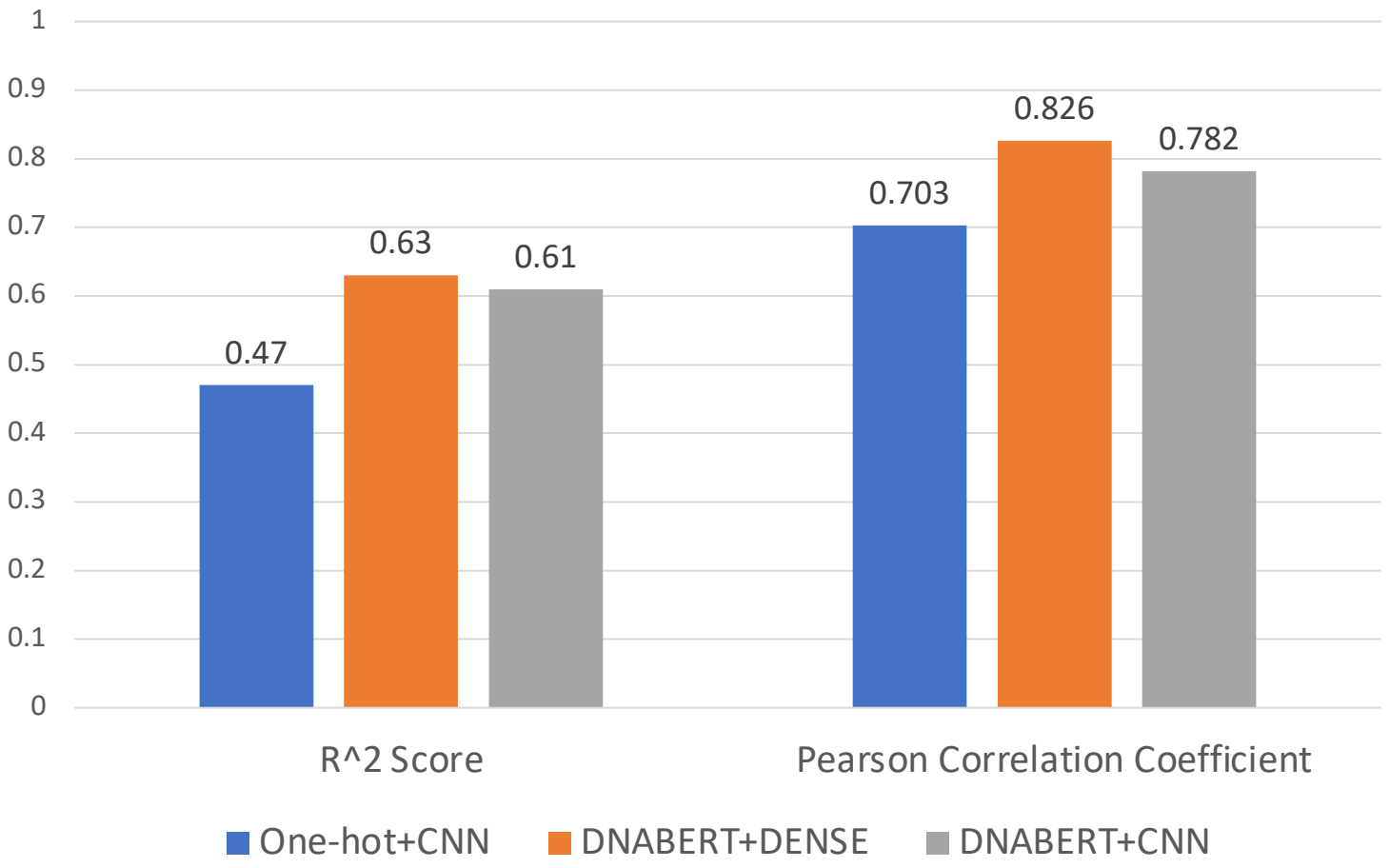
2. DNABERT + Dense



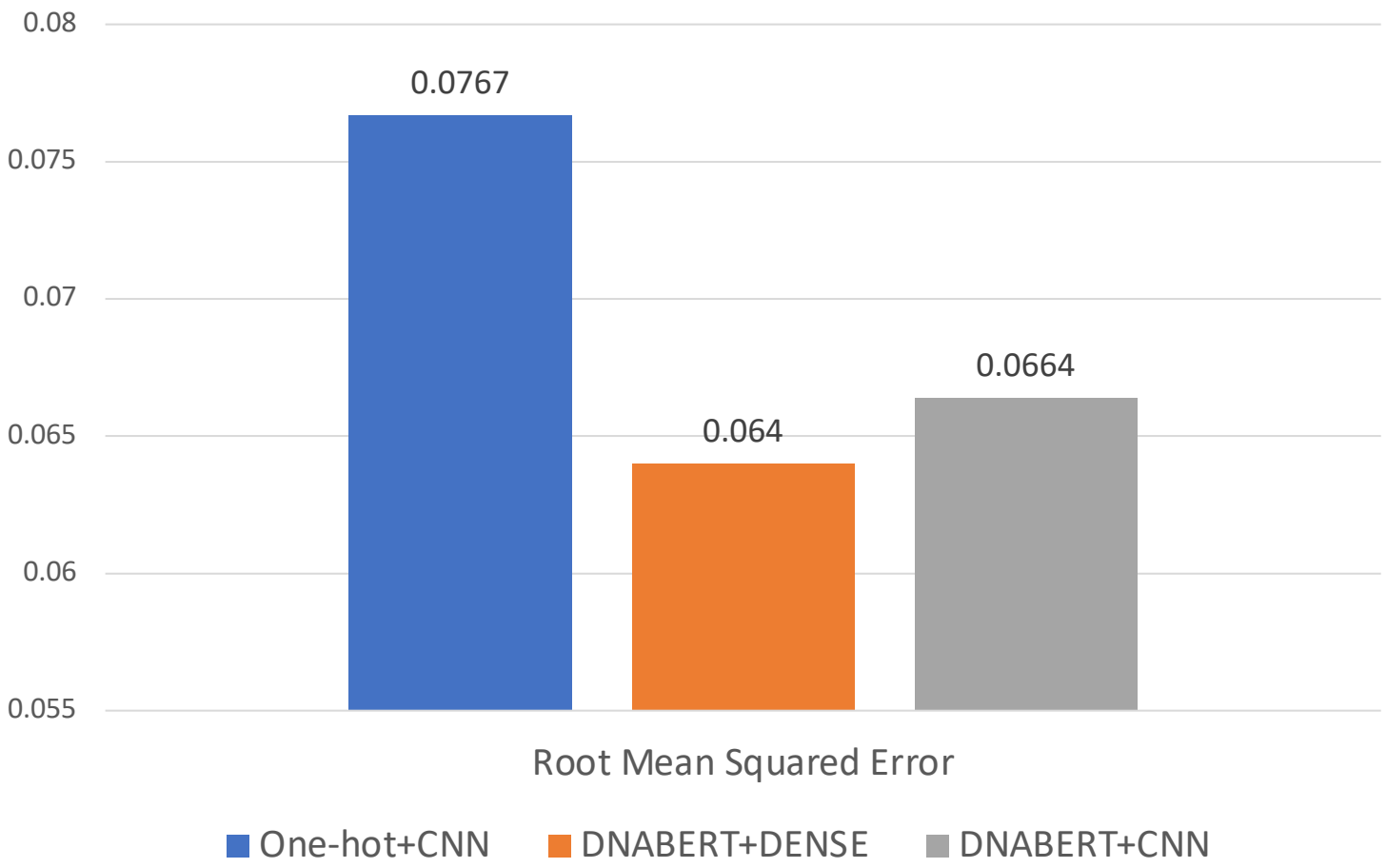
3. DNABERT + CNN



Summary of results



Summary of results



Interpretation

DNABERT generalises well to different tasks and organisms

DNABERT Performs better than one-hot+CNN

CNN is restricted by receptive field, unable to easily take in global contextual information of sequence

Results could imply this is important

Not enough time to spend hyperparameter tuning, these scores could be improved slightly

Future work

Pre-train DNABERT on yeast genome and compare performance

Evaluate the three pipelines on a classification task instead of a regression task

- Need to create a non-promoter dataset

Visualise what DNABERT is focusing its ‘attention’ on

- Could elucidate important features of promoters that make it strong

References

[1]Ayoubi, T.A.Y. and Van De Yen, W.J.M. (1996), Regulation of gene expression by alternative promoters. The FASEB Journal, 10: 453-460. https://doi.org/10.1096/fasebj.10.4.8647344

[2]Oubounyt, M., Louadi, Z., Tayara, H., & Chong, K. T. (2019). DeePromoter: Robust Promoter Predictor Using Deep Learning. Frontiers in genetics, 10, 286. https://doi.org/10.3389/fgene.2019.00286

[3]Kotopka, B.J., Smolke, C.D. Model-driven generation of artificial yeast promoters. Nat Commun 11, 2113 (2020). https://doi.org/10.1038/s41467-020-15977-4

[4]Alammar, J. (2018) The Illustrated Transformer.

http://jalammar.github.io/illustrated-transformer/

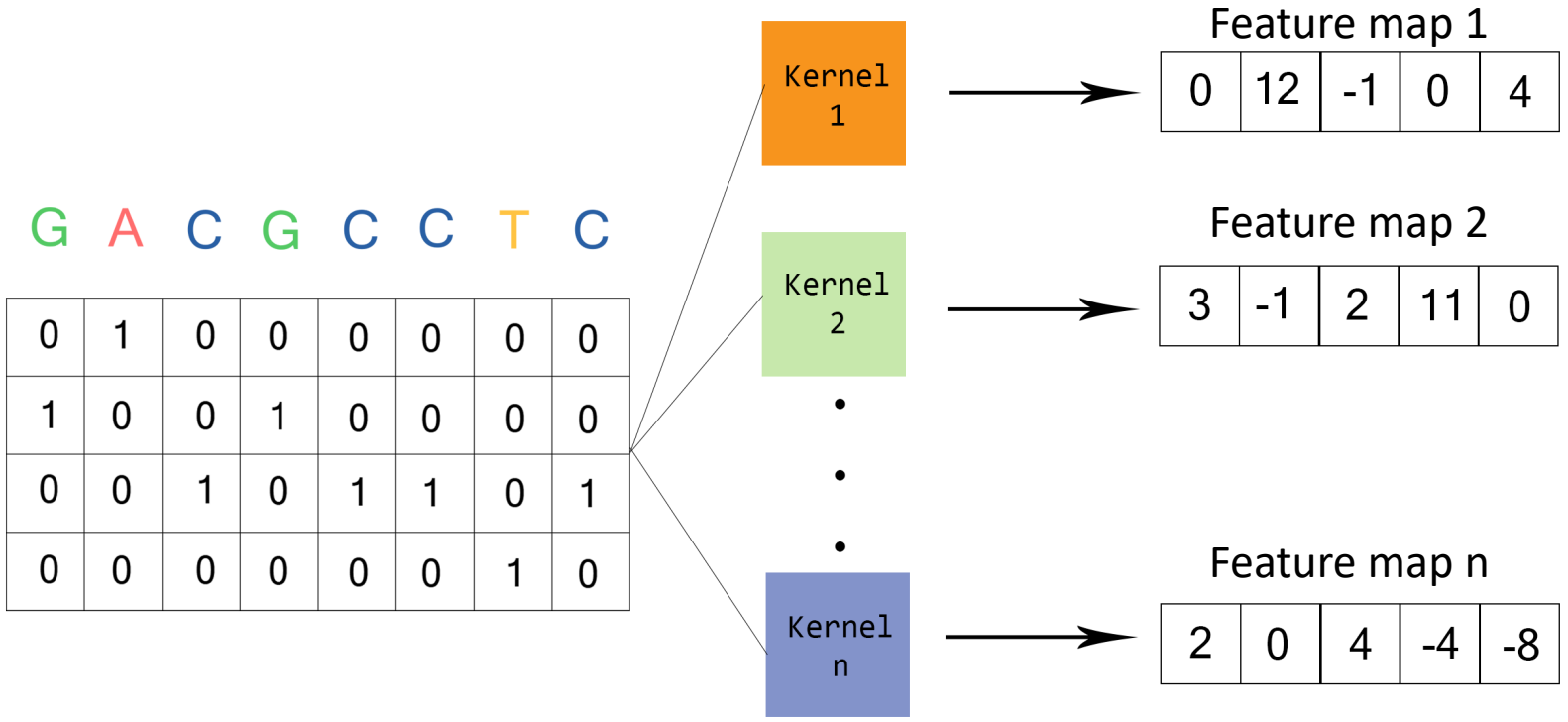
[5]Yanrong Ji, Zhihan Zhou, Han Liu, Ramana V Davuluri, DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome, *Bioinformatics*, 2021, btab083, https://doi.org/10.1093/bioinformatics/btab083

[6]Hugging Face (2020). Summary of the tasks https://huggingface.co/transformers/task_summary.html

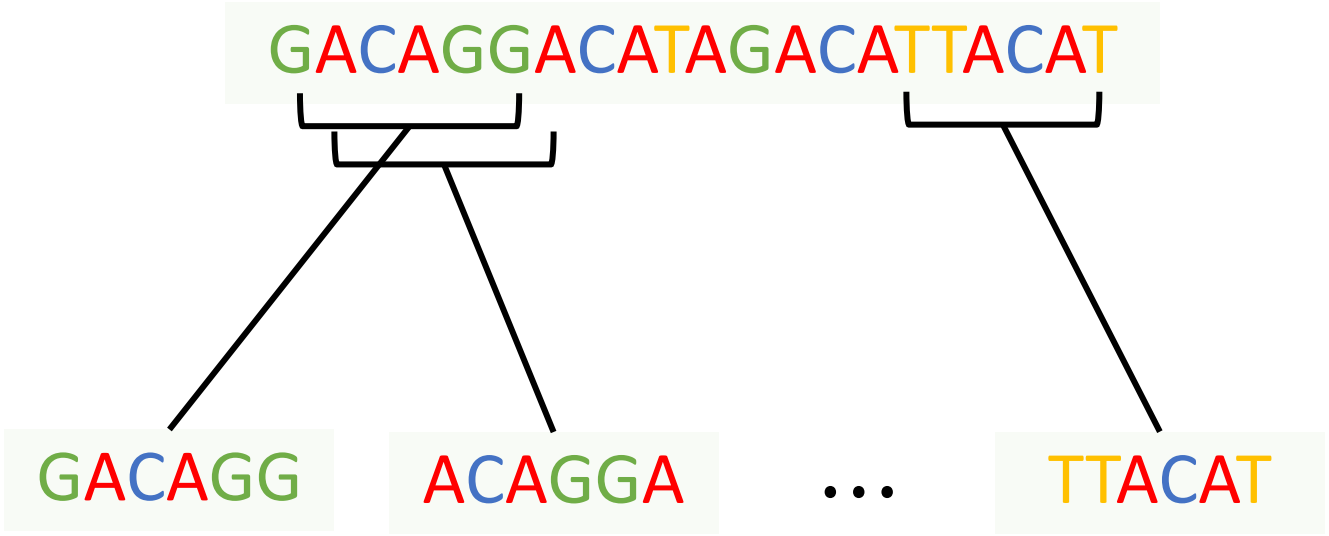
[7]Devlin J. et al. (2018) Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv : 1810.04805*.

Convolutional neural network (CNN)

- Usually create more than one feature map from the same input using different kernels



k-mers



6-mers

4096 possible 6-mers