Computational Systems Biology
Deep Learning in the Life Sciences

6.802 6.874 20.390 20.490 HST.506

Patrick Holec
March 24, 2017

Predicting Enhancer-Promoter Interaction from Genomic Sequence with Deep Neural Networks

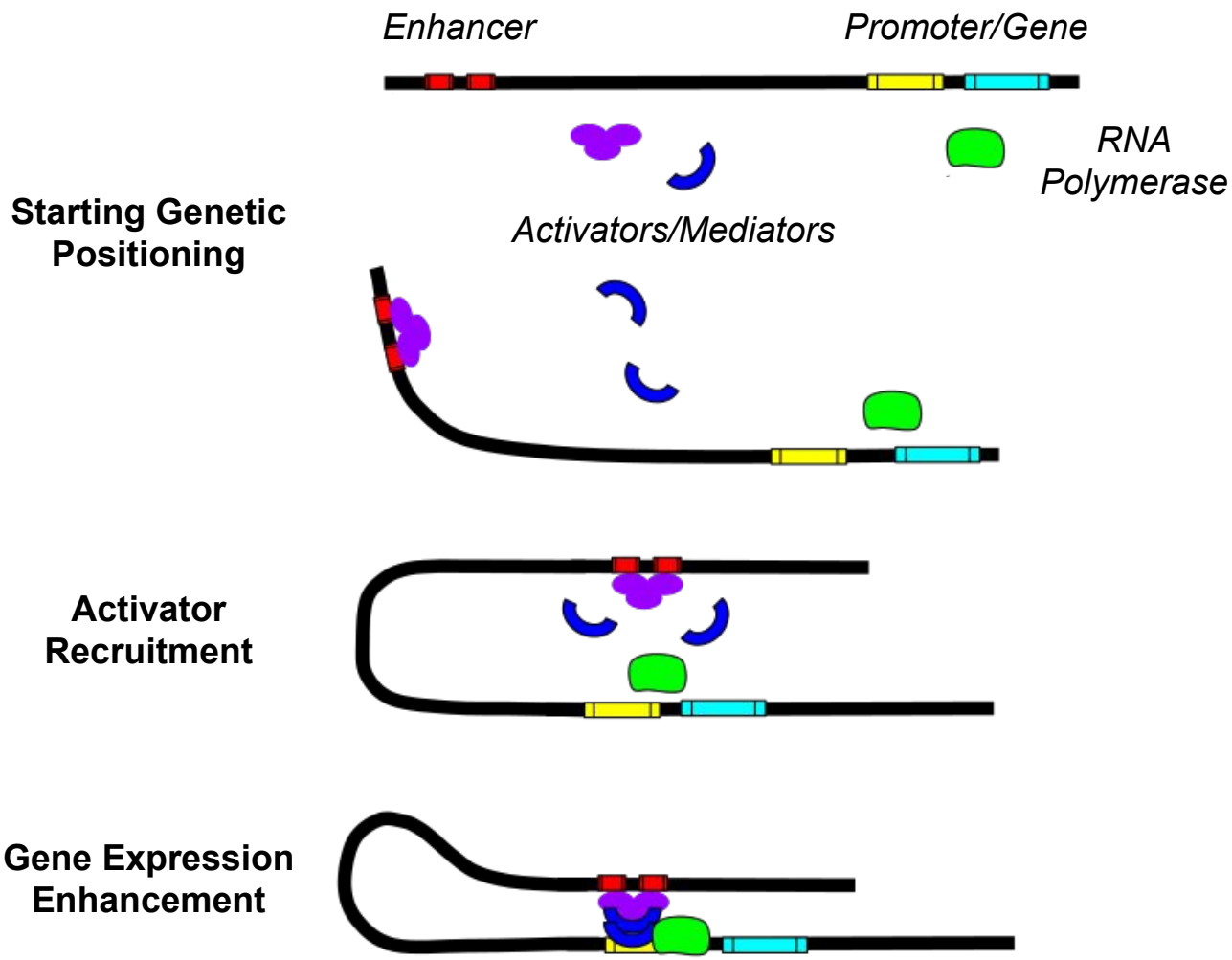
Shashank Singh, Yang Yang, Barnabas Poczos, Jian Ma

http://mit6874.github.io

Key Claim

SPEID is a deep learning pipeline that provides useful predictions for enhancer-promoter interactions

Biological Background



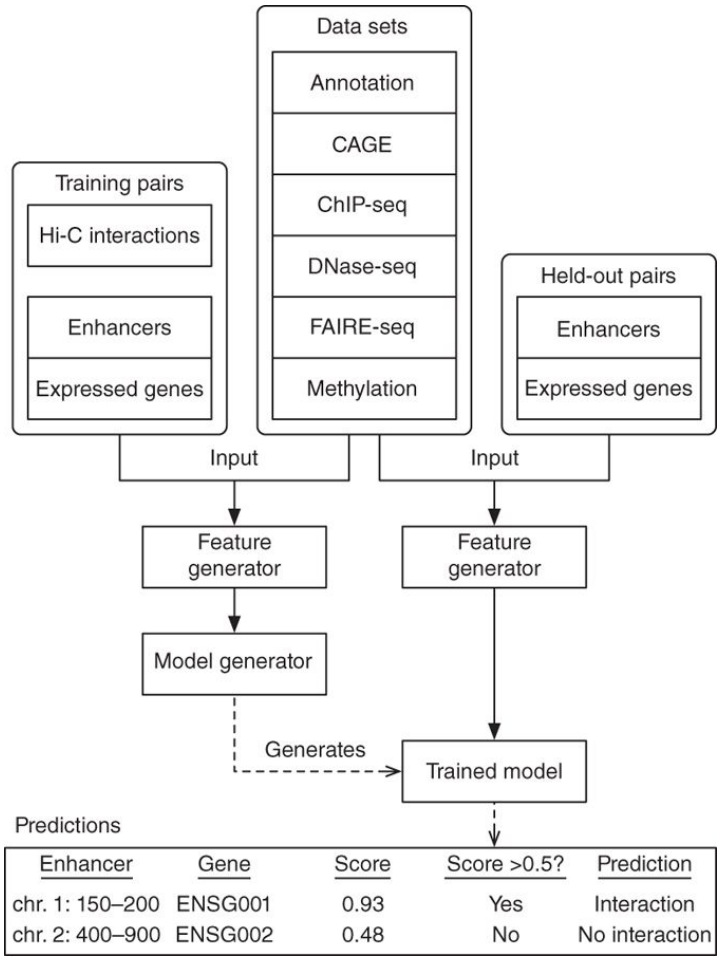
Adapted from Wikipedia

Previous Methods

- Predictions based on sequence features and functional genomic data
- Familiar cross-validation framework for data fitting

TargetFinder

Existing EPI Identification

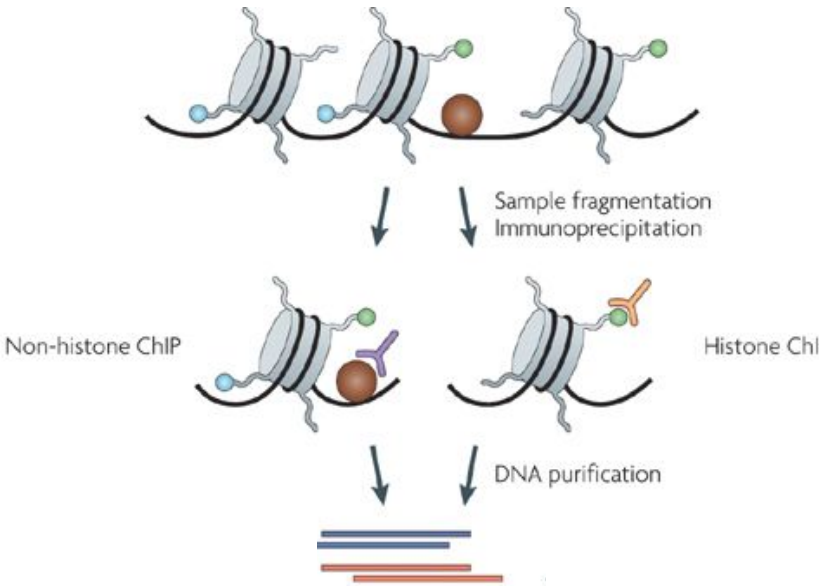


Whalen et al, 2016.

Previous Methods

ChIP-Seq

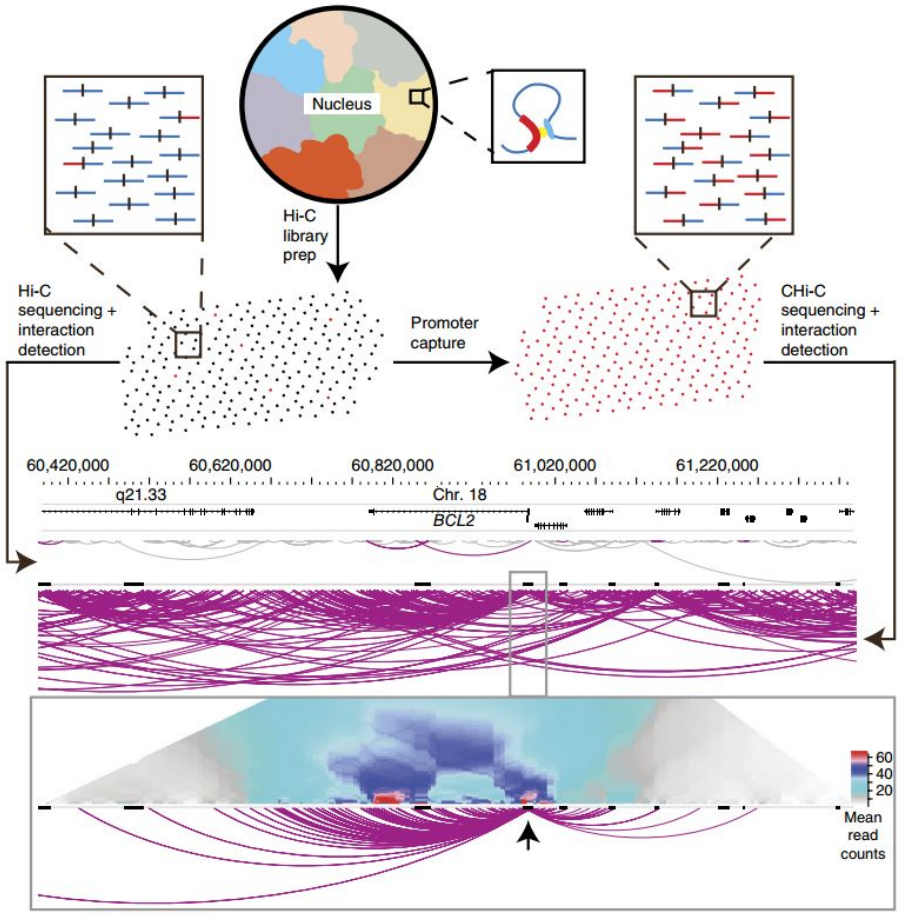
Chromatin/Transcription Factor Mapping



Park et al, 2015.

CHi-C

Enhancer-Promoter Interactions



Mifsud et al, 2015.

Motivation

Q: Can we demonstrate comparable predictive power for EPIs using sequence features exclusively?

Motivation

Q: Can we demonstrate comparable predictive power for EPIs using sequence features exclusively?

A: Yes. Well, kinda.

Assumptions

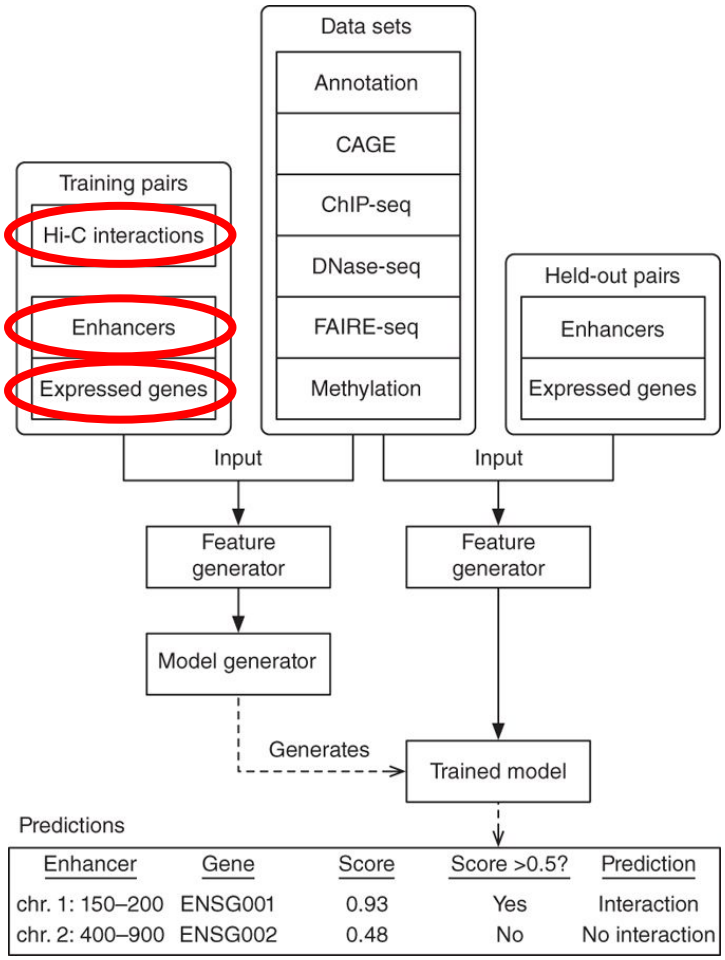
- **The presence of an enhancer-promoter interactions is binary variable**
  - *Model predicts probability of interaction*
- **Sequence features are a defining characteristic to predict relationships**
  - *Two fundamental inputs are sufficient: enhancers and promoters*

Data Used

- All interactions and labels acquired from TargetFinder study for six different cell lines

TargetFinder

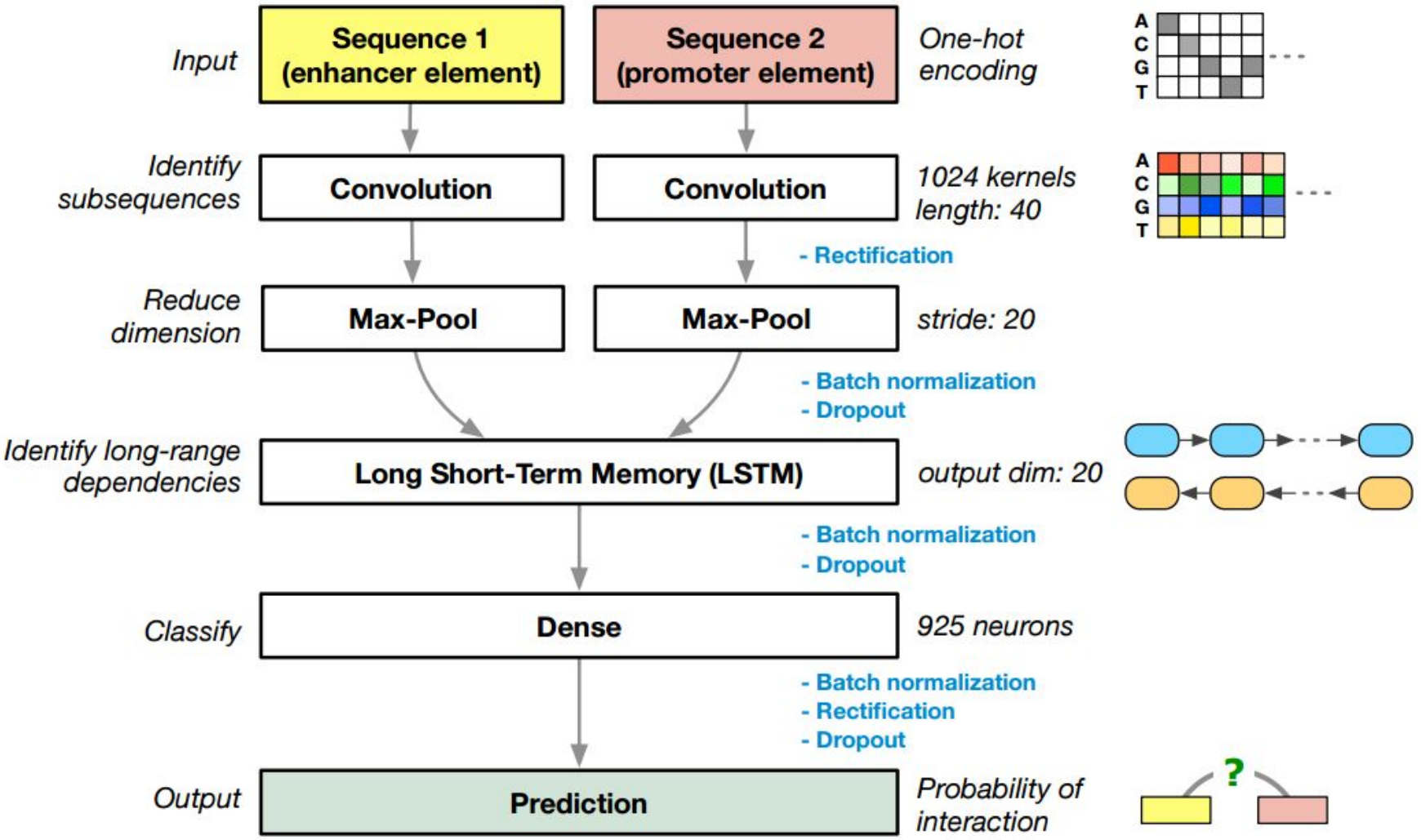
Existing EPI Identification



Whalen et al, 2016.

SPEID Method

Sequence-based Promoter-Enhancer Interaction with Deep Learning



Singh et al, 2016.

LSTM Layer

"The internal mechanism of an LSTM is fairly complex…"
-Singh et al.

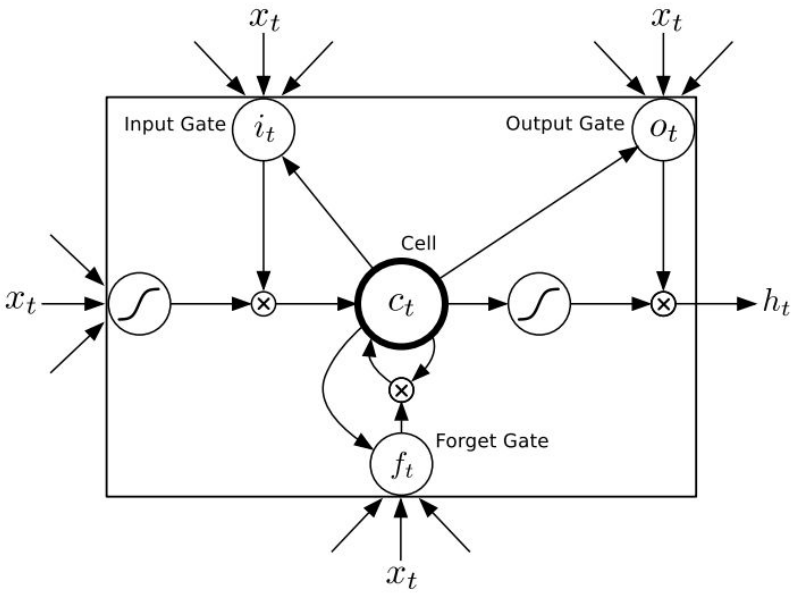


Fig. 1. Long Short-term Memory Cell

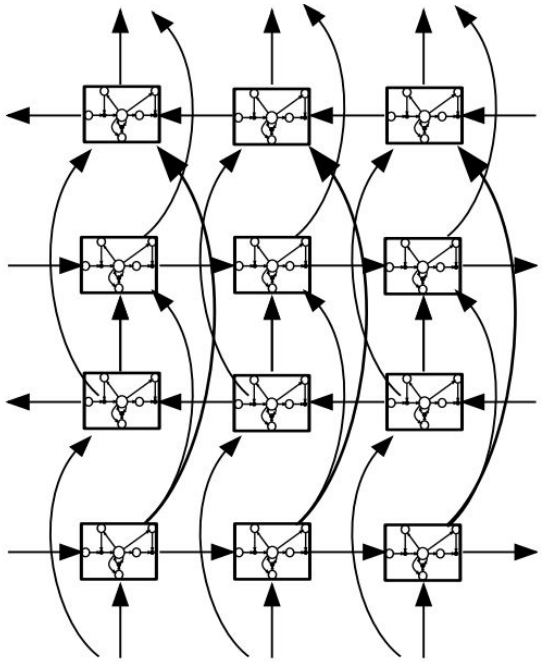
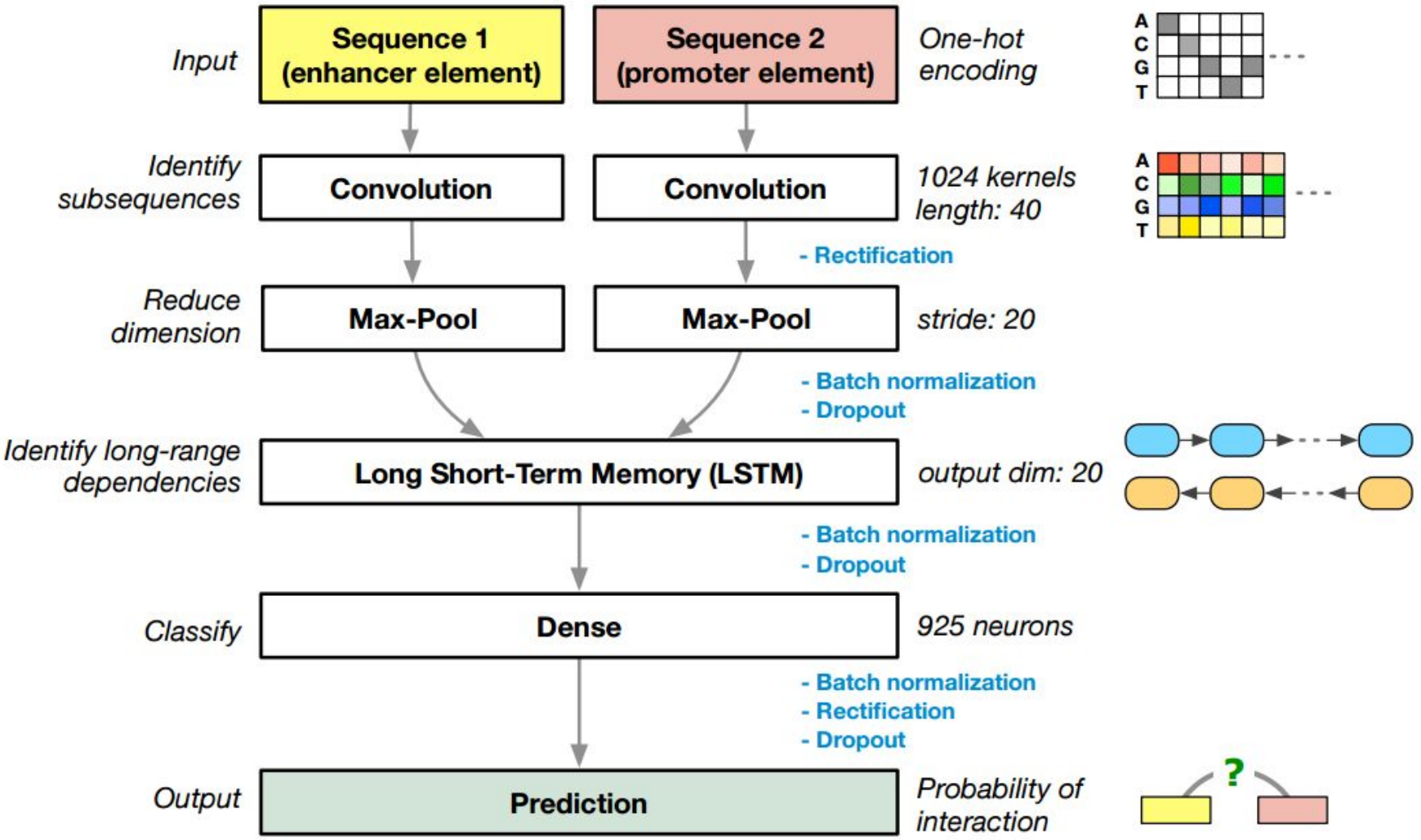


Fig. 4. Deep Bidirectional Long Short-Term Memory Network (DBLSTM)

Graves et al, 2013.

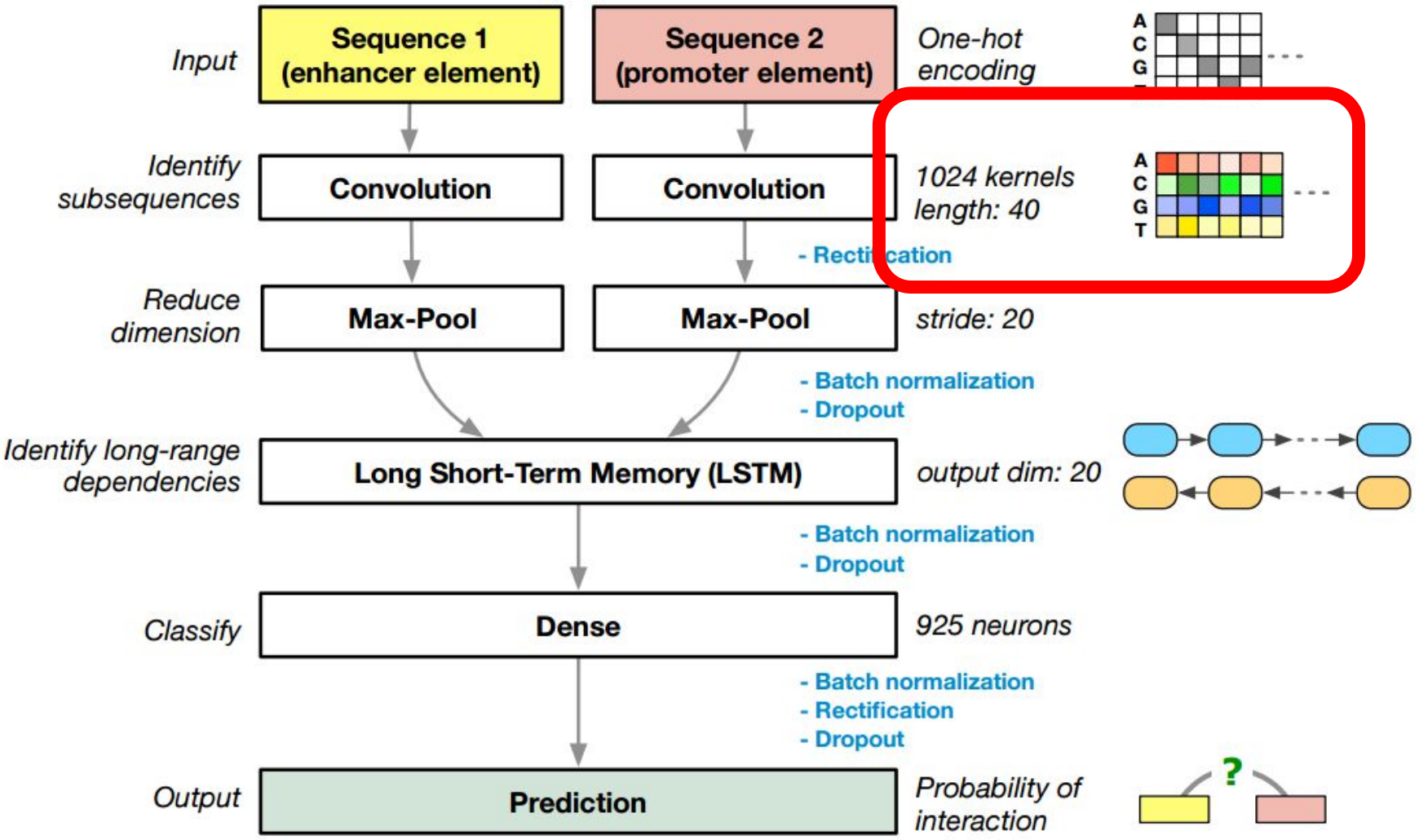
SPEID Method

Sequence-based Promoter-Enhancer Interaction with Deep Learning



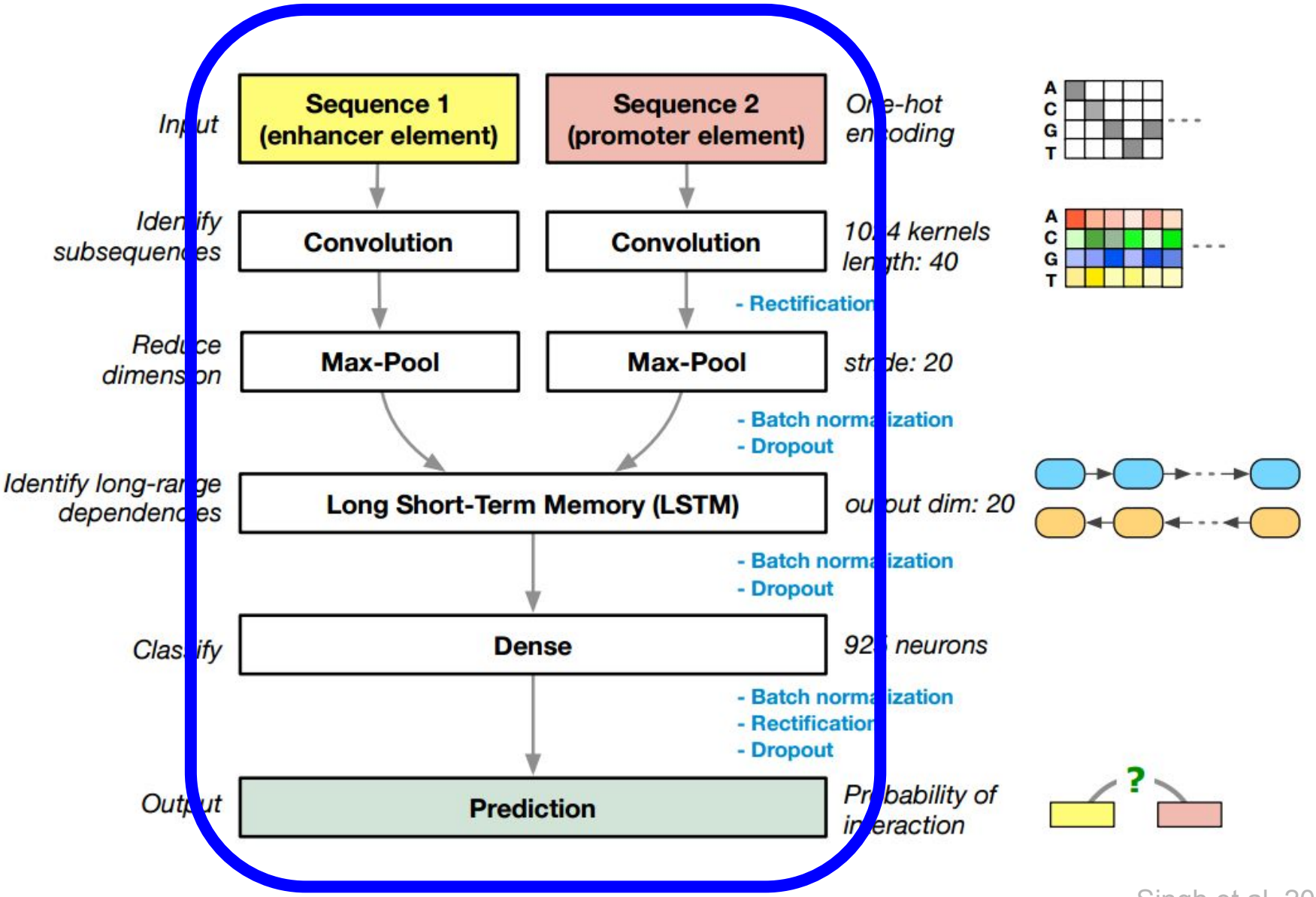
Singh et al, 2016.

Kernels set to JASPAR motifs



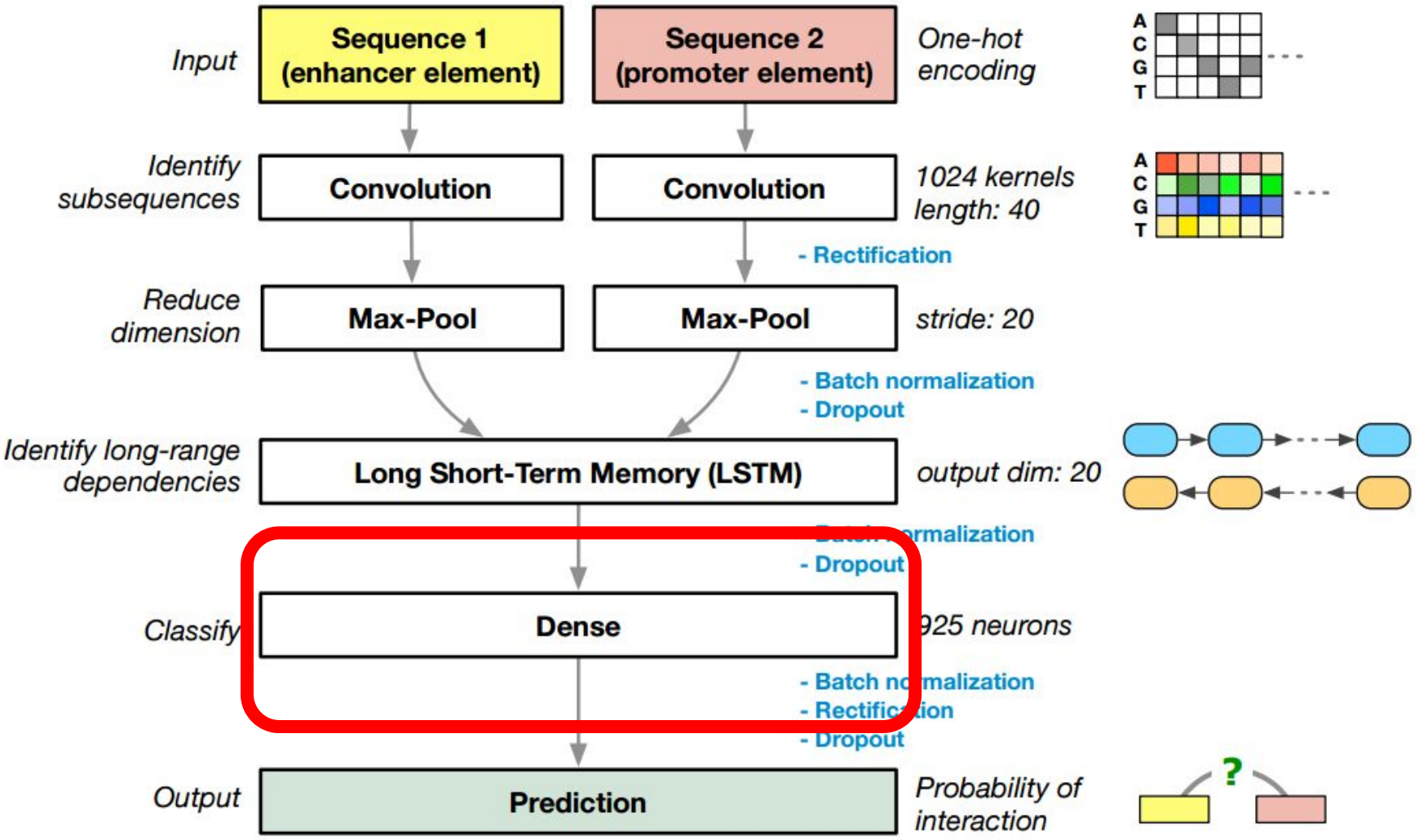
Singh et al, 2016.

Step 1. Train full model



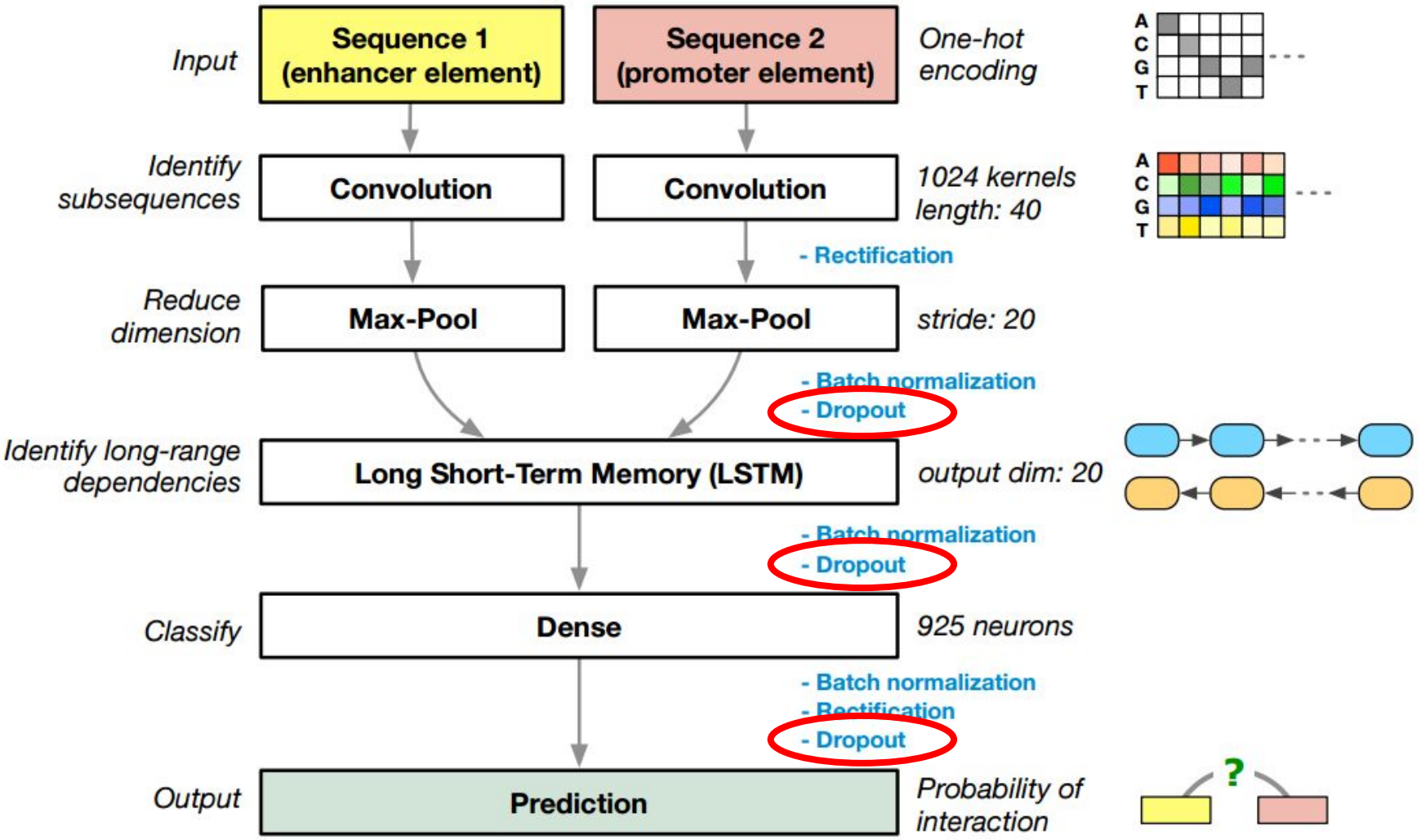
Singh et al, 2016.

Step 2. Train dense layer



Singh et al, 2016.

50% Dropout



Singh et al, 2016.

Results

TargetFinder Benchmark

	Cell Type					
Model	GM12878	HeLa-S3	HUVEC	IMR90	K562	NHEK
SPEID	**0.85**	0.81	0.75	**0.78**	**0.85**	**0.94**
TargetFinder (E/P)	0.59	0.61	0.48	0.48	0.61	0.83
TargetFinder (EE/P)	0.84	0.83	0.71	0.83	0.81	0.83
TargetFinder (E/P/W)	0.81	**0.87**	**0.77**	**0.78**	**0.85**	0.90

Table 2: F_1 scores of different EPI prediction methods for each cell line.

Cell Line Reproducibility

	Testing Cell Type					
Training Cell Type	GM12878	HeLa-S3	HUVEC	IMR90	K562	NHEK
GM12878	**0.87**	0.62	0.64	0.64	0.62	0.59
HeLa-S3	0.60	**0.87**	0.68	0.56	0.62	0.62
HUVEC	0.63	0.67	**0.88**	0.62	0.63	0.66
IMR90	0.62	0.63	0.64	**0.87**	0.60	0.64
K562	0.64	0.63	0.63	0.57	**0.90**	0.59
NHEK	0.58	0.65	0.66	0.56	0.59	**0.88**

Table 4: Area under ROC curve (AUROC) for SPEID when training and testing on different cell lines.

Note: $F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$

Transcription Factor Identification

Matched known transcription factors to kernels identified in SPEID and motifs found in TargetFinder

Cell Line	Predicted important in both	Only in SPEID	Only in TargetFinder
GM12878	22	9	53
HeLa-S3	13	15	37
HUVEC	1	14	7
IMR90	4	31	16
K562	27	26	85
NHEK	0	16	5

Key Claims

- Comparable predictive power to TargetFinder using purely sequence features over functional genomics
- First deep learning model in the enhancer-promoter interaction space, representing a conceptual expansion of the field

Analysis

- ***Very few statistical arguments made***
- Study appears reproducible within their study's framework
- Lack of sensitivity analysis
- ***No numerical model/hyperparameter justification***

Impact

- Shows sequence can be an adequate dataset to define enhancer-promoter interaction motifs (for a particular phenotype)
- ***To do***: further research in additive powers for functional genomic data

Summary

- **Key Claim**
  - *SPEID is a deep learning pipeline that provides useful predictions for enhancer-promoter interactions*
- **Importance**
  - *Improvement on TargetFinder's enhancer-promoter interaction prediction*
- **Issues**
  - *Predictive power is not generalizable across cell lines*
  - *Only mild increase in performance relative to other methods*

Questions?