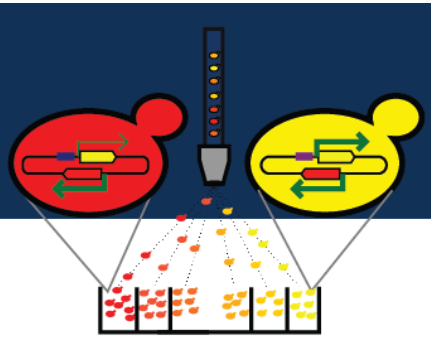
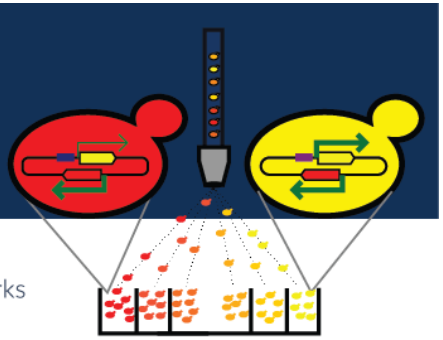


Predicting gene expression using millions of random promoter sequences
DREAM Challenge 2022



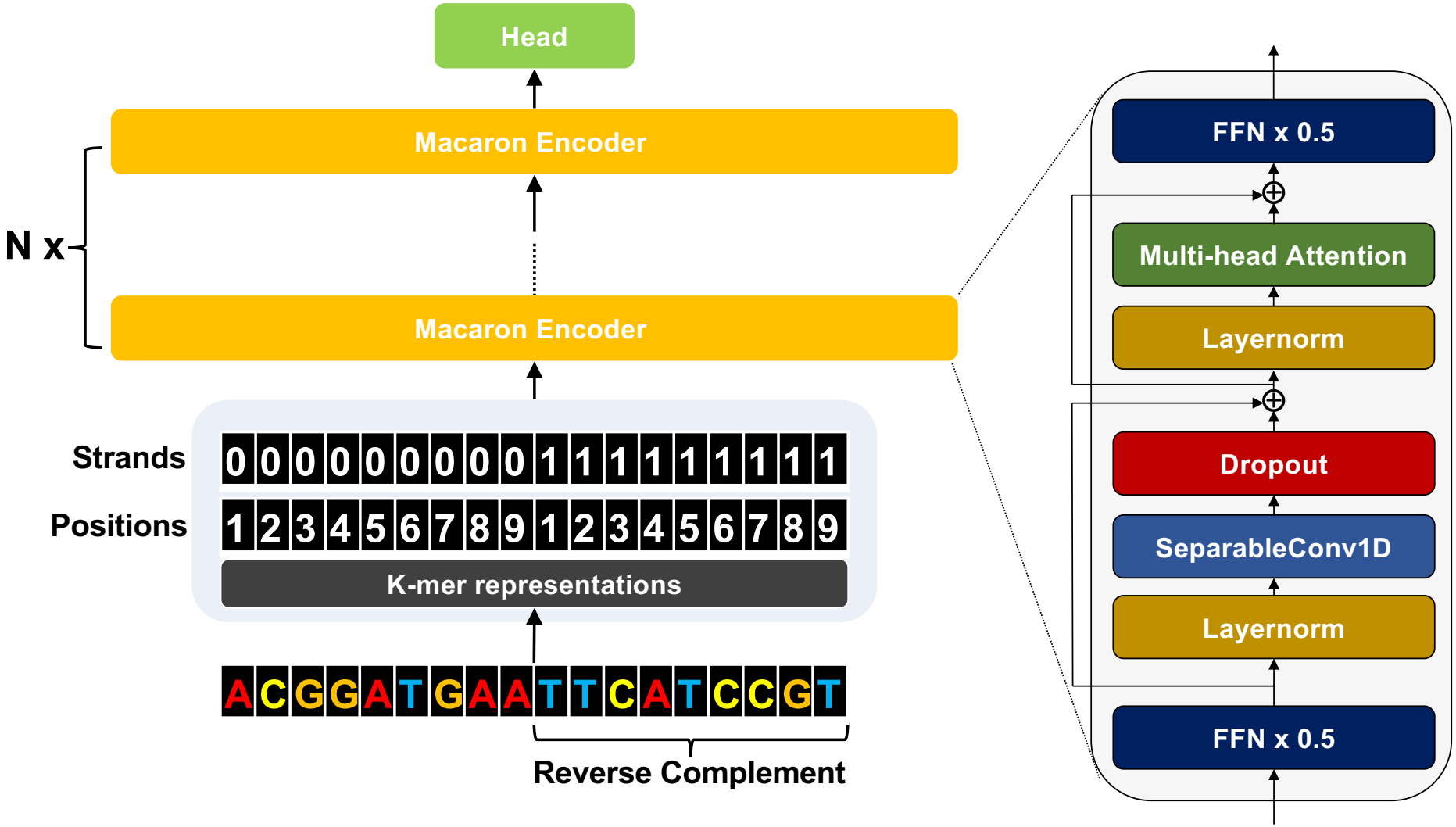
Proformer: a hybrid macaron transformer model predicts expression values from promoter sequences

Wuming Gong[1], Byeong-Chan Kim[2], Juhyun Lee[2], Il-Youp Kwak[2]

[1] Lillehei Heart Institute, University of Minnesota, USA
[2] Department of Applied Statistics, Chung-Ang University, Seoul, Republic of Korea

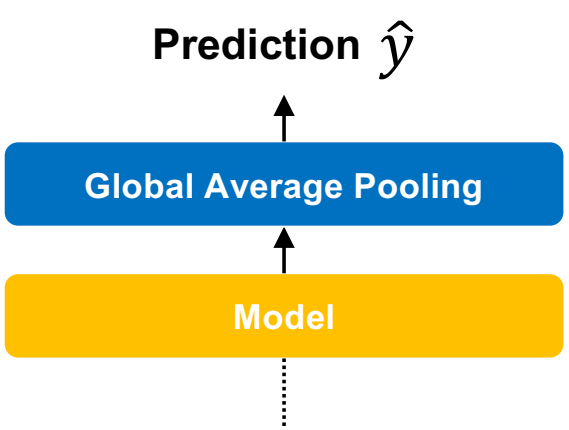
https://github.com/gongx030/dream_PGE

Proformer: an end-to-end Transformer encoder architecture to predict the expression values from DNA sequences

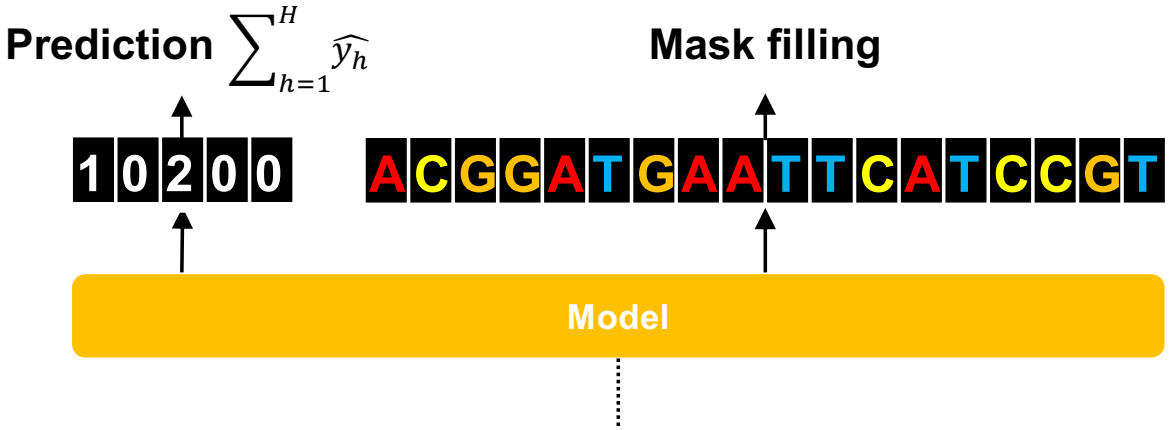


Large over-parameterized models with global average pooling layer failed to converge

Global average pooling

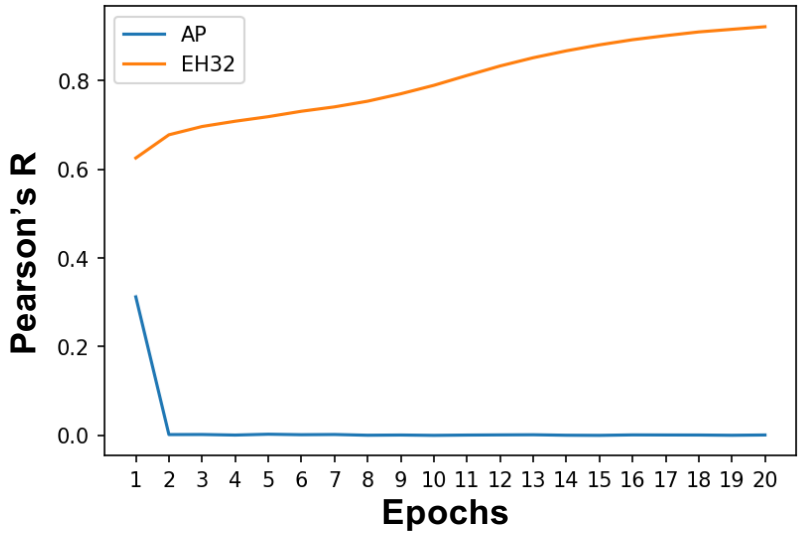


Multiple expression heads with mask filling

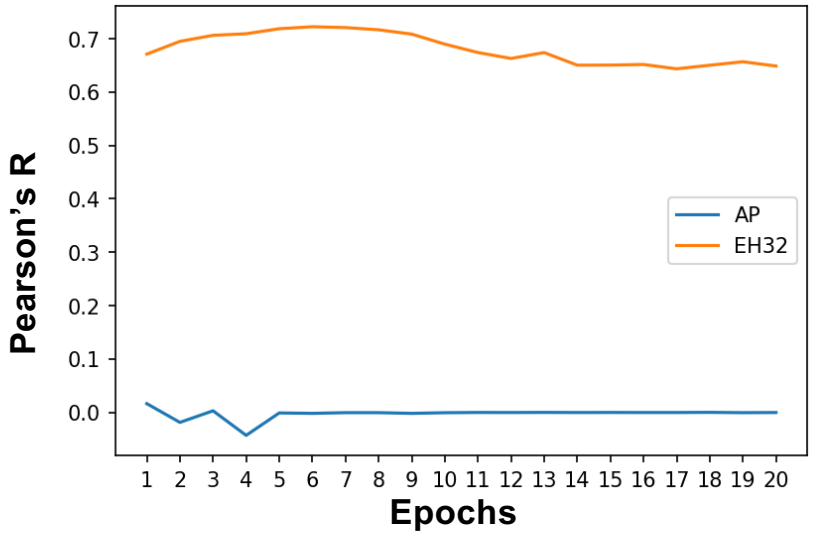


Dimension of 256 and Macaron blocks of 8 on ~500k samples

Training



Validation



Multiple expression heads with mask filling produce better performance on large over-parameterized models

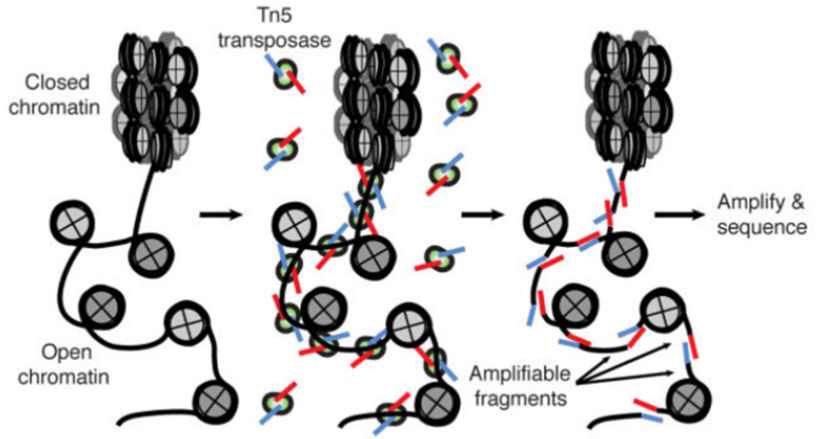
Dim.	Attention heads	Macaron Blocks	Average Pooling	Expression heads (H)				
				1	8	16	32	64
64	2	1	0.7026	**0.7046**	0.7011	0.6977	0.6977	0.6943
64	2	2	0.7094	0.7086	0.7122	**0.7119**	0.7136	0.7088
64	2	4	0.7140	0.7196	0.7162	0.7184	0.7190	**0.7200**
64	2	8	0.7151	0.7198	**0.7223**	0.7138	0.7191	0.7214
128	4	1	0.7033	0.7153	0.7137	0.7075	0.7069	0.7047
128	4	2	0.7164	**0.7209**	0.7142	0.7197	0.7147	0.7175
128	4	4	0.7189	**0.7224**	0.7207	0.7218	0.7139	0.7192
128	4	8	0.0145	0.6627	0.7223	0.7200	0.7207	***0.7226***
256	8	1	0.7109	**0.7177**	0.7104	0.7124	0.7058	0.7152
256	8	2	0.7157	**0.7219**	0.7197	0.7185	0.7207	0.7177
256	8	4	0.6406	0.6616	0.7210	0.7186	**0.7213**	0.7211
256	8	8	0.0165	0.0603	0.7188	0.7194	**0.7222**	0.7173

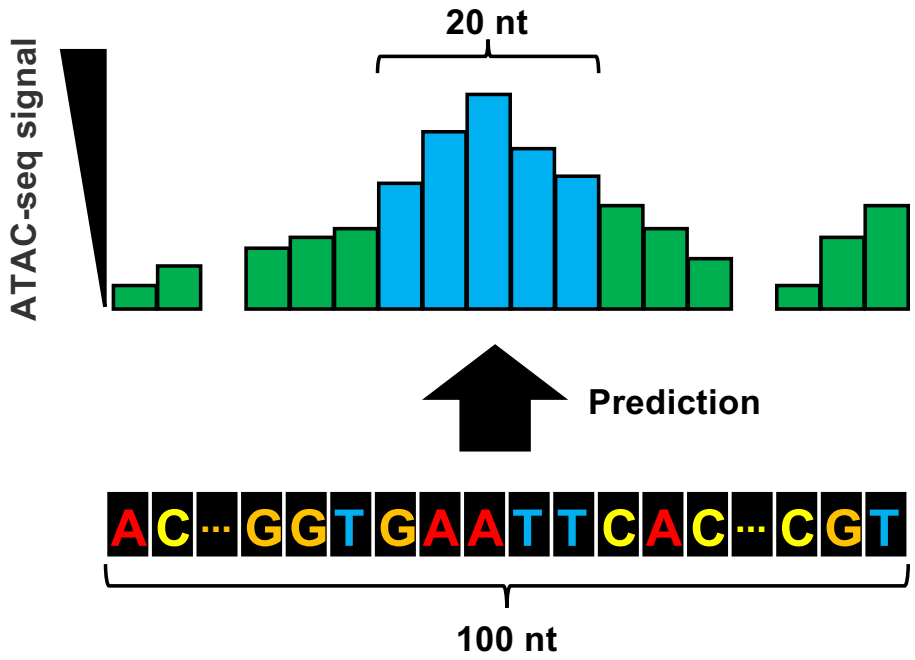
Pearson's R

- 10% of training sequences / expression pairs
- Adam optimizer with base learning rate of 0.001 and clipping
- Linear warmup (one epoch) with cosine decay
- Batch size of 512
- Best Pearson's R in the first 20 epochs
- Masking 5% of positions
- K-mer of 5 with stride of one

Predicting chromatin accessibility from DNA sequences

Assay for Transposase-Accessible Chromatin with high-throughput sequencing (ATAC-seq)





- ATAC-seq data of GM12878 (human LCL)
- 100,000 genomic sub-regions were randomly sampled from the 1,000 bp region surrounding each of ~80,000 ATAC-seq summits.
- Each genomic sub-region includes 100 nucleotides.
- We built models to predict mean ATAC-seq signals of the central 20 bp from DNA sequences.

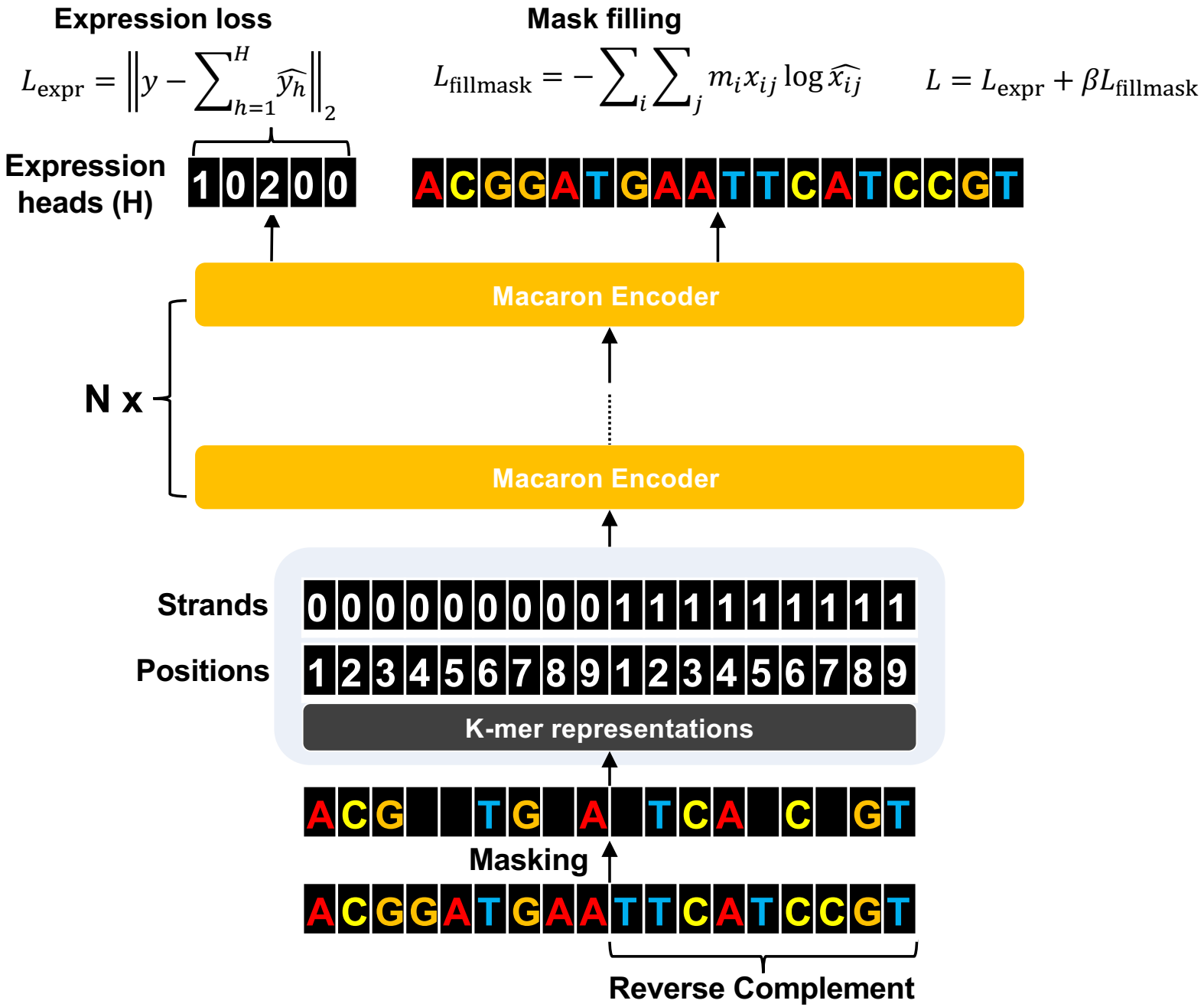
Multiple expression heads with mask filling have better performance on predicting chromatin accessibility from DNA sequences

Dim.	Attention heads	Macaron Blocks	Average Pooling	Expression heads (H)				
				1	8	16	32	64
64	2	1	**0.4726**	0.4497	0.4450	0.4570	0.4467	0.4528
64	2	2	**0.4832**	0.4353	0.4739	0.4677	0.4660	0.4626
64	2	4	0.4434	**0.4871**	0.4823	0.4855	0.4834	0.4783
64	2	8	0.4222	**0.4888**	0.4767	0.4828	0.4875	0.4848
128	4	1	0.4434	0.4451	**0.4660**	0.4651	0.4627	0.4574
128	4	2	0.4177	0.4711	**0.4889**	0.4802	0.4847	0.4850
128	4	4	0.2346	0.3977	***0.4964***	0.4868	0.4880	0.4882
128	4	8	0.0155	0.0858	0.0335	**0.4946**	0.4910	0.4915

Pearson's R

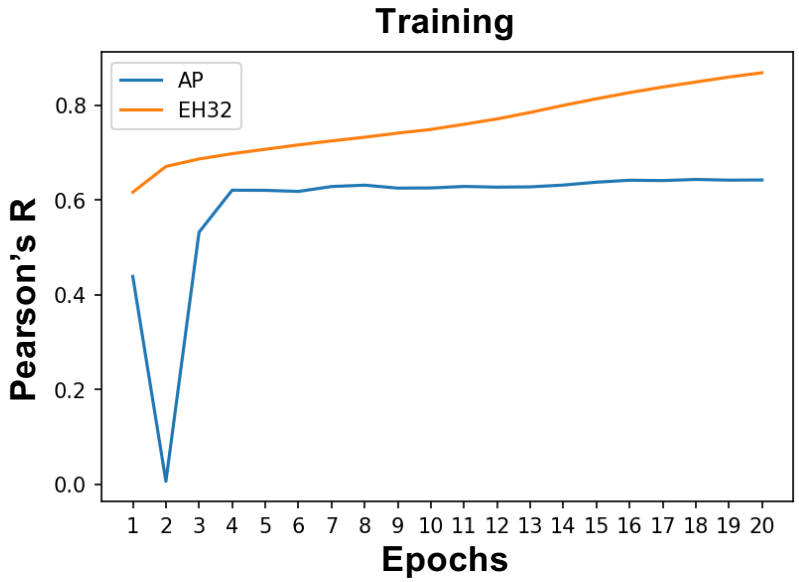
- 80k training samples
- 20k testing samples
- Adam optimizer with base learning rate of 0.001 and clipping
- Linear warmup (one epoch) with cosine decay
- Batch size of 512
- Best Pearson's R in the first 20 epochs
- Masking 5% of positions
- K-mer of 5 with stride of one

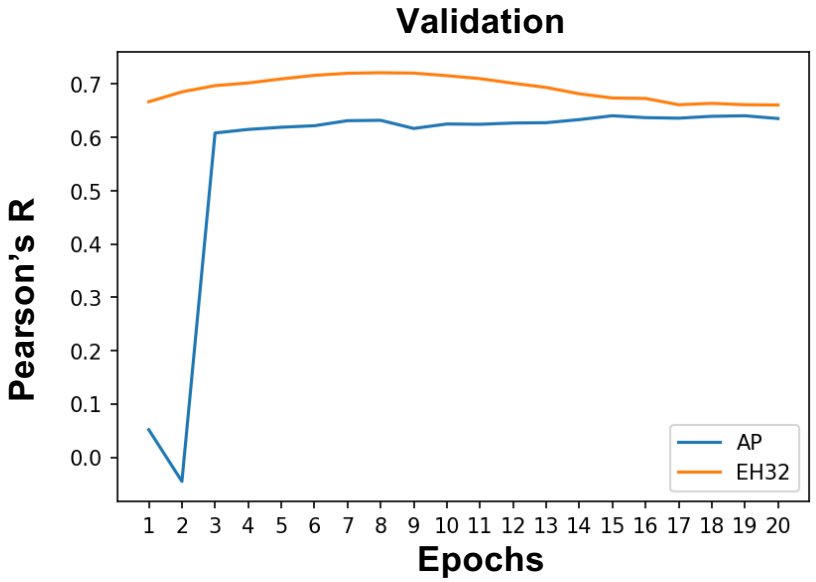
Proformer: A hybrid Macaron transformer model predicts expression values from promoter sequences



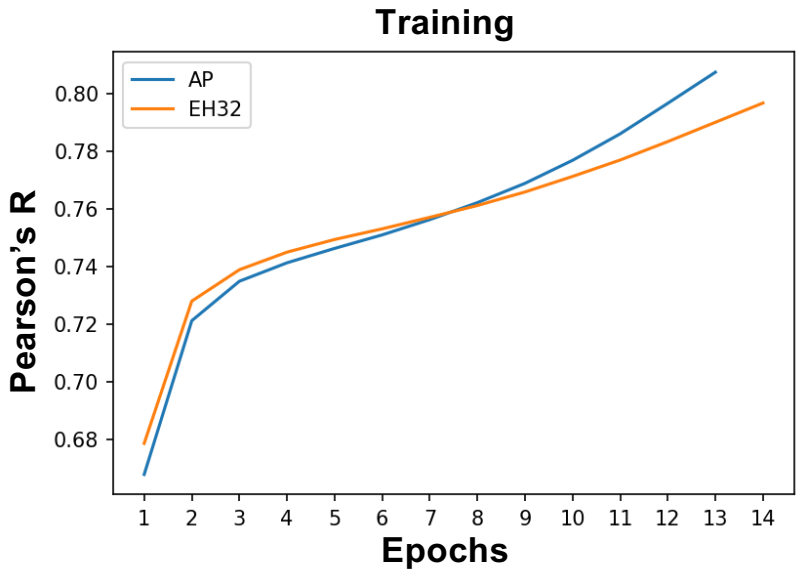
Multiple expression heads with mask filling produce better performance on larger models

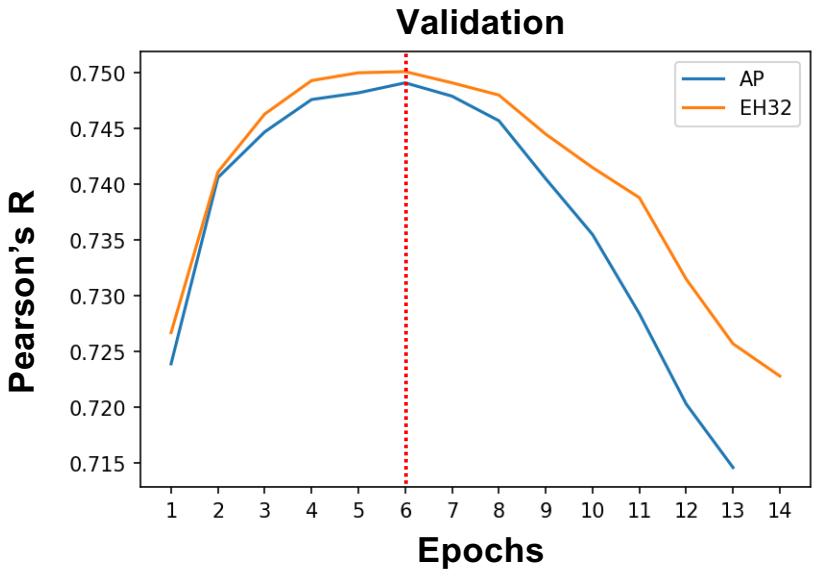
10% data





full data



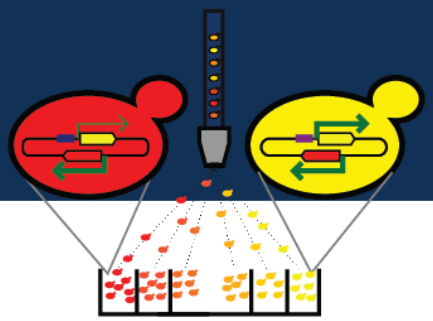


Multiple expression heads with mask filling are critical for improving the prediction performance on hold-out validation data

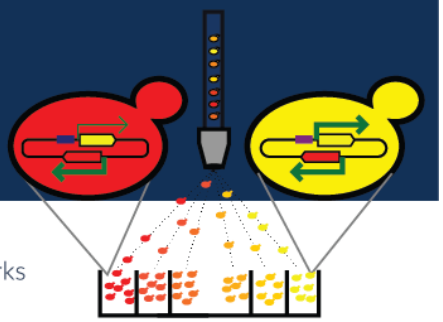
Ablation analysis

Global average pooling	Mask filling	Expr. heads	Score PearsonR	Score Spearman	PearsonR	Spearman
X			0.766	0.819	0.918	0.961
	X	1	0.765	0.817	0.921	0.964
	X	32	**0.781**	**0.827**	0.926	0.965
	X	32*	0.765	0.810	**0.929**	**0.967**

* with GLU



Predicting gene expression using millions of random promoter sequences
DREAM Challenge 2022



SBME UBC School of Biomedical Engineering

IBM Research

Google Research TPU Research Cloud

SageBionetworks

Challenge Organizers

- Carl de Boer
- Abdul Muntakim Rafi
- Jake Albrecht
- Pablo Meyer
- Paul Boutros

All participants

CAU ET lab

- Byeong-Chan Kim
- Juhyun Lee
- Il-Youp Kwak

Lillehei Heart Institute

- Nikita Dsouza
- Xiao Ma
- Daniel J Garry

We thank Minnesota Supercomputing Institute, University of Minnesota, Driven to Discover® for providing computation resource.

We thank deep genomics for providing travel grant of $2,000.