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

Il-Youp Kwak[1], Byeong-Chan Kim[1], Juhyun Lee[1], Daniel J. Garry[2,3,4*], Jianyi Zhang[5], and Wuming Gong[1*],

Affiliations:

[1]Department of Applied Statistics, Chung-Ang University, Seoul, Republic of Korea

[2]Cardiovascular Division, Department of Medicine, University of Minnesota, Minneapolis, MN 55455

[3]Stem Cell Institute, University of Minnesota, Minneapolis, MN 55455

[4]Paul and Sheila Wellstone Muscular Dystrophy Center, University of Minnesota, Minneapolis, MN 55455

[5]Department of Biomedical Engineering, The University of Alabama at Birmingham, Birmingham, AL 35233

*Co-corresponding authors:

Wuming Gong, Ph.D.

Lillehei Heart Institute

2231 6th St SE

University of Minnesota

Minneapolis, MN 55455

Phone: 612-298-1881

Email: gongx030@umn.edu

Daniel J. Garry, M.D., Ph.D.

Lillehei Heart Institute

2231 6th St SE

University of Minnesota
Minneapolis, MN 55455
Phone: 612-626-2178
Email: garry@umn.edu

Abstract

The breakthrough high-throughput measurement of the cis-regulatory activity of millions of randomly generated promoters provides an unprecedented opportunity to systematically decode the cis-regulatory logic that determines the expression values. We developed an end-to-end transformer encoder architecture named Proformer to predict the expression values from DNA sequences. Proformer used a Macaron-like Transformer encoder architecture, where two half-step feed forward (FFN) layers were placed at the beginning and the end of each encoder block, and a separable 1D convolution layer was inserted after the first FFN layer and in front of the multi-head attention layer. The sliding *k*-mers from one-hot encoded sequences were mapped onto a continuous embedding, combined with the learned positional embedding and strand embedding (forward strand vs. reverse complemented strand) as the sequence input. Moreover, Proformer introduced multiple expression heads with mask filling to prevent the transformer models from collapsing when training on relatively small amount of data. We empirically determined that this design had significantly better performance than the conventional design such as using the global pooling layer as the output layer for the regression task. These analyses support the notion that Proformer provides a novel method of learning and enhances our understanding of how cis-regulatory sequences determine the expression values.

Introduction

Gene expression is a fundamental process and is essential for the coordinated function of all living organisms. Predicting the expression level of a gene based on its promoter or enhancer sequences is an important problem in molecular biology, with applications ranging from understanding the regulation of gene expression to engineering gene expression for biotechnological applications[1,2]. Recent progress and mechanistic insights have been obtained using large-scale and high-throughput massively parallel reporter assays (MPRAs), which enable the study of gene expression and regulatory elements in a high-throughput manner and the simultaneous testing of thousands to millions of enhancers or promoters in parallel[3–24]. MPRA protocols linked random or mutated sequences to unique barcodes, with each sequence-barcode pair represented in a different reporter assay vector. After delivery of the pooled vector library, barcode abundance could be subsequently quantified using next-generation sequencing (NGS) techniques[25]. MPARs enabled large scale studies of functional annotation of putative regulatory elements[3,26], variant effect prediction[22,23,27,28] and evolutionary reconstructions[25,29,30]. For example, STARR-seq (self-transcribing active regulatory region sequencing) was used to investigate the enhancer activities of tens of millions of independent fragments from the Drosophila genome[3]. Microarray-based or PCA-based (polymerase cycling assembly) synthesized DNA regulatory elements with unique sequence tags were used to evaluate hundreds of thousands of variants of mammalian promoters or enhancers[4–6]. Nguyen et al. systematically compared the promoter and enhancer potentials of many candidate sequences[10]. Using Gigantic Parallel Reporter Assay (GPRA), de Boer et al. measured the expression level associated with tens of millions of random promoter sequences and used these to learn cis-regulatory logic in the yeast grown in well-characterized carbon sources[14].

Machine learning methods have been developed to identify complex relationships and patterns in large scale DNA sequences (including MPRA data) that may not be apparent through conventional statistical methods. For example, convolutional neural networks (CNN) and recurrent neural networks (RNN) were used to capture the local

dependences in DNA sequences and/or genomic features and predict binding affinities[1,31,32], chromatin features[33,34], DNA methylation[35,36], RBP (RNA-binding protein) binding[37–39] and gene expression levels[40]. In contrast, Transformers are a type of neural network architecture that has gained popularity in recent years for their ability to process sequential data, such as text and speech, more efficiently and effectively than traditional RNNs and CNNs[41]. Transformers used an attention mechanism to selectively focus on different aspects of the input sequence, which allowed them to capture long-range dependencies more effectively than RNNs and CNNs that typically rely on fixed-length windows or sliding windows.

In this study, we developed an end-to-end transformer encoder architecture, Proformer, to predict the expression values from millions of DNA sequences. Our method introduces several innovative designs such as Macaron-like encoder structures, *k*-mer embedding, and multiple expression heads (MEH) to learn the relationships between a large number of sequences and expression values. Proformer ranked in the 3rd place in the final standing of the DREAM challenge: predicting gene expression using millions of random promoter sequences[42]. We believe that our model provides a novel method of learning and characterizing how cis-regulatory sequences determine the expression values. Codes pertaining to important analyses in this study are available from GitHub webpage: https://github.com/gongx030/dream_PGE.

Results

Proformer overview

Proformer used a Macaron-like Transformer encoder architecture to predict the expression values from promoter sequences (Figure 1) [43–45]. Compared with the regular Transformer encoder, the Macaron-like encoder has two half-step feed forward (FFN) layers at the beginning and the end of each Transformer encoder block, which can be mathematically interpreted as a numerical Ordinary Differential Equation (ODE) solver for a convection-diffusion equation in a multi-particle dynamic system [46,47]. Given the stochastic nature of the input sequences, we hypothesized that this design may better recover the associations between nucleotide pattern and the expression values. We added a separable 1D convolution layer in the Macaron encoder block following the first FFN layer and in front of the multi-head attention layer. This design has been used in other Transformer architectures such as Conformer[47], and is shown to be critical for capturing the local signals.

We extracted the sliding *k*-mers (*k*=10 in the final model) from one-hot encoded sequences and mapped them onto a continuous embedding platform. It has been previously shown that the *k*-mer embedding of nucleotide sequences had better performance than the convolution on tasks such as predicting transcription factor binding sites[48]. The *k*-mer embedding was then combined with the learned positional embedding and strand embedding (forward strand vs reverse complemented strand) as one part of the input to the Macaron encoder.

We added H positions ($H = 32$ in our final model) as the expression heads (Figure 1). Proformer predicted one expression value for each expression head and used the mean of the prediction of all positions as the final predicted expression value. The total training losses consisted of the mean squared error between predicted and observed expression values ($\boldsymbol{L}_{expr}$), and the reconstruction loss ($\boldsymbol{L}_{recon}$), where we randomly

masked 5% of the nucleotides and had the model predict the masked nucleotides. In our final model, we set the weight for reconstruction loss $\beta = 1$.

The final Proformer model had approximately 47 million trainable parameters, implemented by TensorFlow 2 and trained on one machine with four A100 GPUs. We varied the learning rate over the course of training according to the formula used in the original Transformer paper[43]. Warmup steps of 12,500 and a batch size of 512 were used in the training. We used the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$ and $\epsilon = 10^{-9}$ for these studies.

MEH with mask filling has improved performance using large over-parameterized models

Global average pooling layer at the top of a neural network is commonly used for the regression and classification tasks[49]. However, we found that when applying the global average pooling layer at the top of a large transformer model, for example, with a dimension size of 256 and blocks size of 8, the whole model sometimes failed to converge on training on relatively small amount (~500k) of samples (Figure 2b). In order to address this issue, we proposed a new design, where the model predicted multiple expression values through multiple expression heads (MEH) and used the average of all predictions as the final predicted value (Figure 2a), while at the same time, the model also predicted the randomly masked DNA nucleotide. MEH with mask filling produced stable convergence when training the transformer model with the same size on ~500k samples (Figure 2b). In order to systematically compare the performance of two designs, we trained the models on 10% of the training sequence / expression value pairs then the performance was evaluated on 2% of the data as the Pearson's R between observed and predicted expression values. For MEH with mask filling, we also examined the performance over a different number of heads ($H = 1,8,16,32,64$). Overall, we found that MEH with mask filing gave significantly better than global average pooling when using 8 or more heads (Mann-Whitney U test p-values = 0.0715, 0.0102, 0.0142, 0.0224 and 0.00605 for $H = 1,8,16,32,64$, respectively), and the best performance was

achieved at a dimension size of 128 and macaron block size of 8 (Figure 2c). As the model size became larger and deeper, the global average pooling became difficult to converge, while in comparison, MEH with masking filing could still provide stable results.

MEH with mask filling has better performance for the prediction of chromatin accessibility from DNA sequences

To test whether our observations on these two head designs could apply to similar scenarios, we designed another task to use Proformer to predict ATAC-seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) signals from DNA sequences. The ATAC-seq is a technique to measure the chromatin accessibility across the whole genome[50]. We sampled a total of 100k genomic sub-regions surrounding the ~80,000 summits of ATAC-seq data of GM12878[50], while each genomic sub-region included 100 nucleotides. Different models were built to predict the mean ATAC-seq signal of the central 20 bp from 100 nt DNA sequences (Figure 3a). The global average pooling performed well when the model size was relatively small. As the model size became larger, we observed similar trends such that the global average pooling tended to fail on large over-parameterized models. The best performance was achieved by using MEH with mask filling with dimension size of 128 and block size of 4 (Figure 3b).

MEH with mask filling is critical for improving the prediction performance on hold-out validation data

We trained the final model for the DREAM challenge by using a dimension size of 512 and a block size of 4 on 95% of the data provided by the organizers and evaluated on the remaining 5%. The checkpoint after the 6th epoch was used where the validation Pearson's R was maximized. As expected, MEH with mask filling produced improved Pearson's R than global average pooling on the validation data (Figure 4a). The ablation study showed when using only one expression head ($H = 1$), the performance was similar to global average pooling. However, MEH with $H = 32$ showed improvement

over hold-out validation data and produced the highest weighted scores. It is interesting that adding a GLU activation[51] to expression heads produced even higher unweighted Pearson's R and Spearman's Rho on the hold-out validation data, while the weighted score became worse than global average pooling (Figure 4b). Future studies will explore different designs of the expression heads.

Discussion

Various machine learning techniques have been used to analyze and interpret the MPRA data and dissect the regulatory logics. Recently, over-parameterized deep networks or large models, with more parameters than the size of the training data, have dominated the performance in various machine learning areas[52]. The global average pooling layer was conventionally used to aggregate the information from multiple channels and to produce final predictions. However, we found that when training over-parameterized models on the regression tasks such as predicting expression values from DNA sequences, the global average pooling often led to a convergence issue, most likely due to the loss of information that accumulated during the training and caused the model to perform poorly or failed to converge. Here we presented a new architecture Proformer for prediction of expression values from DNA sequences. We introduced a new design named multiple expression heads (MEH) with mask filling to prevent the over-parameterized transformer models from collapsing when training on relatively small amount of data. Applying the Proformer model to predictexpression values and to predict chromatin accessibility from DNA sequences showed that MEH with masking filling produced significantly better performance and stable convergence compared to the commonly used global average pooling. Based on our studies, we propose that MEH with mask filling will be a useful design for similar regression tasks that took advantage of large over-parameterized models.

Methods

DREAM challenge dataset overview

Rafi et al. conducted a high-throughput experiment to measure the regulatory effect of millions of random DNA sequences. They cloned 80 bp random DNA sequences into a promoter-like context upstream of a yellow fluorescent protein (YFP), transformed the resulting library into yeast, and measured expression by fluorescent activated cell sorting[4,14,53]. The training dataset includes 6,739,258 random promoter sequences and their corresponding mean expression values[42].

Rafi et al. also provided 71,103 sequences from several promoter sequence types as the hold-out "validation" dataset to evaluate the model performance in different ways. These validation datasets included predicting the expression changes resulting from single nucleotide variants (SNVs), perturbation of specific transcription factor (TF) binding sites, tiling of TF binding sites across background sequences, sequences with high- and low-expression levels, native yeast genomic sequences, random DNA sequences, and challenging sequences designed to maximize differences between a convolutional model and a biochemical model trained on the same data[42].

Sequence trimming and padding

We removed the leading 17 and trailing 13 nucleotides (nt) that were identical in both training and testing promoter sequences, since these nucleotides were not informative for the prediction of expression values and removal of the nucleotides would significantly reduce the training and inference time. The length of the resulting promoter sequences ranged from 48 to 112 nt for training data, while >99.97% training promoters were less than 100 nucleotides. To further reduce the computational overhead, we used 6,737,568 promoter sequences shorter than 100 nt (after trimming) in the model training. For promoters that were less than 100 nt, the left and right sides were padded with the letter N.

Reverse complemented sequences

We empirically found that including the reverse complemented promoter sequences would significantly improve the performance. Thus, the reverse complemented sequences were concatenated with the original sequences (after trimming and padding) and used as the input for model training. Thus, the total length of the input sequences was 200 nt.

Standardization of the expression values

The expression values were standardized to the mean of zero and standard deviation of one. Our experiments found that when the mean squared error loss was used, standardizing the expression values gave better model generalization performance (in terms of Pearson's R and Spearman's Rho) and faster convergence.

ATAC-seq data from GM12878

The human EBV-transformed lymphoblastoid cell line (LCL) ATAC-seq data were downloaded from NCBI GEO database (GSE47753). The sequence reads from three replicates of 50k cell sample (GSM1155957, GSM1155958 and GSM1155959) were pooled and used for the downstream analysis. The 86,004 peaks called by MACS2(v2.1.1)[54,55] were used for the downstream analysis.

Acknowledgement

These studies were supported by funding from NHLBI (P01HL160476), Department of Defense (W81XWH2110606) and Minnesota Regenerative Medicine. We acknowledge the Minnesota Supercomputing Institute for providing computational resources.

286 **Figure 1.** Proformer is a macaron-like transformer architecture that models the
287 relationship between DNA sequences and expression values.
288

289 **Figure 2. Multiple expression heads (MEH) with mask filling has better**
290 **performance on large over-parameterized models. (a)** Global average pooling layer
291 and MEH with mask filling were used at the top the transformer blocks. **(b)** The training
292 (left) and validation (left) performance of Proformer models using global average pooling
293 (AP) or MEH with 32 heads (EH32) were compared. The performance was measured
294 by the Pearson's R between observed and predicted expression values. **(c)** Systematic
295 evaluation of global average pooling and MEH with mask filing on different model
296 specifications such as dimension heads (2, 4, and 8), macaron blocks (1, 2, 4, and 8),
297 and number of expression heads (1, 8, 16, 32, 64) was performed. The best
298 performance of each model specification was highlighted.
299
300

Figure 3. Multiple expression heads (MEH) with mask filling has better performance on predicting chromatin accessibility from DNA sequences. (a) The task of predicting mean ATAC-seq signal of the central 20 bp from 100 nt surrounding DNA sequences was examined. **(b)** Systematic evaluation of global average pooling and MEH with mask filing on different model specifications such as dimension heads (2 and 4), macaron blocks (1, 2, 4, and 8), and number of expression heads (1, 8, 16, 32, 64). The best performance of each model specification was highlighted.

Figure 4. Multiple expression heads (MEH) with mask filling is critical for improving the prediction performance on hold-out validation data. (a) The training (left) and validation (left) performance of Proformer models on the full DREAM dataset using global average pooling (AP) or MEH with 32 heads (EH32). The performance was measured by the Pearson's R between observed and predicted expression values. The checkpoint after the 6th epoch was used as the final model where the validation Pearson's R was maximized (red dotted line). **(b)** The performance of Proformer model on the hold-out validation data. The performance is measured by weighted (score) or unweighted Pearson's R and Spearman's Rho between observed and predicted expression values.

References

1. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat Biotechnol* 1–9 (2015) doi:10.1038/nbt.3300.

2. Bussemaker, H. J., Foat, B. C. & Ward, L. D. Predictive Modeling of Genome-Wide mRNA Expression: From Modules to Molecules. *Annu Rev Bioph Biom* 36, 329–347 (2007).

3. Arnold, C. D. *et al.* Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science* 339, 1074–1077 (2013).

4. Sharon, E. *et al.* Inferring gene regulatory logic from high-throughput measurements of thousands of systematically designed promoters. *Nat Biotechnol* 30, 521–530 (2012).

5. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat Biotechnol* 30, 271–277 (2012).

6. Patwardhan, R. P. *et al.* Massively parallel functional dissection of mammalian enhancers in vivo. *Nat Biotechnol* 30, 265–270 (2012).

7. Kheradpour, P. *et al.* Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. 23, 800–811 (2013).

8. Lubliner, S. *et al.* Core promoter sequence in yeast is a major determinant of expression level. *Genome Res* 25, 1008–1017 (2015).

9. Farley, E. K. *et al.* Suboptimization of developmental enhancers. *Science* 350, 325–328 (2015).

10. Nguyen, T. A. *et al.* High-throughput functional comparison of promoter and enhancer activities. *Genome Res* 26, 1023–1033 (2016).

11. Smith, R. P. *et al.* Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nat Genet* 45, 1021–1028 (2013).

12. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the in vivo effects of binding site variants. *Genome Res* 23, 1908–1915 (2013).

13. Arensbergen, J. van *et al.* Genome-wide mapping of autonomous promoter activity in human cells. *Nat Biotechnol* 35, 145–153 (2017).

354 14. Boer, C. G. de *et al.* Deciphering eukaryotic gene-regulatory logic with 100 million
355 random promoters. *Nat Biotechnol* 38, 56–65 (2020).

356 15. Gertz, J., Siggia, E. D. & Cohen, B. A. Analysis of combinatorial cis-regulation in
357 synthetic and genomic promoters. *Nature* 457, 215–218 (2009).

358 16. Weingarten-Gabbay, S. *et al.* Systematic interrogation of human promoters.
359 *Genome Res* 29, 171–183 (2019).

360 17. Grossman, S. R. *et al.* Systematic dissection of genomic features determining
361 transcription factor binding and enhancer function. *Proc National Acad Sci* 114, E1291–
362 E1300 (2017).

363 18. Shen, S. Q. *et al.* Massively parallel cis-regulatory analysis in the mammalian
364 central nervous system. *Genome Res* 26, 238–255 (2016).

365 19. Haberle, V. *et al.* Transcriptional cofactors display specificity for distinct types of
366 core promoters. *Nature* 570, 122–126 (2019).

367 20. Klein, J. C. *et al.* A systematic evaluation of the design and context dependencies of
368 massively parallel reporter assays. *Nat Methods* 17, 1083–1091 (2020).

369 21. Vockley, C. M. *et al.* Massively parallel quantification of the regulatory effects of
370 noncoding genetic variation in a human cohort. *Genome Res* 25, 1206–1214 (2015).

371 22. Tewhey, R. *et al.* Direct Identification of Hundreds of Expression-Modulating
372 Variants using a Multiplexed Reporter Assay. *Cell* 165, 1519–1529 (2016).

373 23. Ulirsch, J. C. *et al.* Systematic Functional Dissection of Common Genetic Variation
374 Affecting Red Blood Cell Traits. *Cell* 165, 1530–1545 (2016).

375 24. Liu, S. *et al.* Systematic identification of regulatory variants associated with cancer
376 risk. *Genome Biol* 18, 194 (2017).

377 25. Romero, I. G. & Lea, A. J. Leveraging massively parallel reporter assays for
378 evolutionary questions. *Genome Biol* 24, 26 (2023).

379 26. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput
380 functional testing of ENCODE segmentation predictions. *Genome Res* 24, 1595–1602
381 (2014).

382 27. Castaldi, P. J. *et al.* Identification of Functional Variants in the FAM13A Chronic
383 Obstructive Pulmonary Disease Genome-Wide Association Study Locus by Massively
384 Parallel Reporter Assays. *Am J Resp Crit Care* 199, 52–61 (2018).

385 28. Shen, S. Q. *et al.* A candidate causal variant underlying both enhanced cognitive
386 performance and increased risk of bipolar disorder. *Biorxiv* 580258 (2021)
387 doi:10.1101/580258.

388 29. Klein, J. C., Keith, A., Agarwal, V., Durham, T. & Shendure, J. Functional
389 characterization of enhancer evolution in the primate lineage. *Genome Biol* 19, 99
390 (2018).

391 30. Arnold, C. D. *et al.* Quantitative genome-wide enhancer activity maps for five
392 Drosophila species show functional enhancer conservation and turnover during cis-
393 regulatory evolution. *Nat Genet* 46, 685–692 (2014).

394 31. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep
395 learning–based sequence model. *Nat Methods* 12, 931–934 (2015).

396 32. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the
397 accessible genome with deep convolutional neural networks. 26, 990–999 (2016).

398 33. Kelley, D. R. *et al.* Sequential regulatory activity prediction across chromosomes
399 with convolutional neural networks. *Genome Res* 28, 739–750 (2018).

400 34. Dsouza, N., Gong, W. & Garry, D. J. SeATAC: a tool for exploring the chromatin
401 landscape and the role of pioneer factors. *Biorxiv* 2022.04.25.489439 (2022)
402 doi:10.1101/2022.04.25.489439.

403 35. Zeng, H. & Gifford, D. K. Predicting the impact of non-coding variants on DNA
404 methylation. *Nucleic Acids Res* 45, gkx177 (2017).

405 36. Angermueller, C., Lee, H. J., Reik, W. & Stegle, O. DeepCpG: accurate prediction of
406 single-cell DNA methylation states using deep learning. *Genome Biol* 18, 67 (2017).

407 37. Pan, X., Rijnbeek, P., Yan, J. & Shen, H.-B. Prediction of RNA-protein sequence
408 and structure binding preferences using deep convolutional and recurrent neural
409 networks. *Bmc Genomics* 19, 511 (2018).

410 38. Budach, S. & Marsico, A. pysster: classification of biological sequences by learning
411 sequence and structure motifs with convolutional neural networks. *Bioinformatics* 34,
412 3035–3037 (2018).

413 39. Avsec, Ž., Barekatain, M., Cheng, J. & Gagneur, J. Modeling positional effects of
414 regulatory sequences with spline transformations increases prediction accuracy of deep
415 neural networks. *Bioinformatics* 34, 1261–1269 (2018).

416 40. Zhou, J. *et al.* Deep learning sequence-based ab initio prediction of variant effects
417 on expression and disease risk. *Nat Genet* 50, 1171–1179 (2018).

418 41. Vaswani, A. *et al.* Attention Is All You Need. *Arxiv* (2017)
419 doi:10.48550/arxiv.1706.03762.

420 42. Rafi, A. M. Evaluation and optimization of sequence-based gene regulatory deep
421 learning models. (2023).

422 43. Vaswani, A. *et al.* Attention Is All You Need. *arXiv.org* cs.CL, (2017).

423 44. Press, O., Smith, N. A. & Levy, O. Improving Transformer Models by Reordering
424 their Sublayers. *Arxiv* (2019).

425 45. Lu, Y. *et al.* Understanding and Improving Transformer From a Multi-Particle
426 Dynamic System Point of View. *Arxiv* (2019).

427 46. Lu, Y. *et al.* Understanding and Improving Transformer From a Multi-Particle
428 Dynamic System Point of View. *Arxiv* (2019) doi:10.48550/arxiv.1906.02762.

429 47. Gulati, A. *et al.* Conformer: Convolution-augmented Transformer for Speech
430 Recognition. *Arxiv* (2020).

431 48. Shen, Z., Bao, W. & Huang, D.-S. Recurrent Neural Network for Predicting
432 Transcription Factor Binding Sites. *Scientific reports* 8, 15270 (2018).

433 49. Pak, M. & Kim, S. A Review of Deep Learning in Image Recognition. *2017 4th Int*
434 *Conf Comput Appl Information Process Technology Caipt* 1–3 (2017)
435 doi:10.1109/caipt.2017.8320684.

436 50. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J.
437 Transposition of native chromatin for fast and sensitive epigenomic profiling of open
438 chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 1213–1218
439 (2013).

440 51. Dauphin, Y. N., Fan, A., Auli, M. & Grangier, D. Language Modeling with Gated
441 Convolutional Networks. *Arxiv* (2016).

442 52. Liu, S., Zhu, Z., Qu, Q. & You, C. Robust Training under Label Noise by Over-
443 parameterization. *Arxiv* (2022) doi:10.48550/arxiv.2202.14026.

444 53. Kinney, J. B., Murugan, A., Callan, C. G. & Cox, E. C. Using deep sequencing to
445 characterize the biophysical mechanism of a transcriptional regulatory sequence. *Proc*
446 *National Acad Sci* 107, 9158–9163 (2010).

447 54. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9, R137
448 (2008).

449 55. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment
450 using MACS. *Nature protocols* 7, 1728–1740 (2012).

451

Figure 1

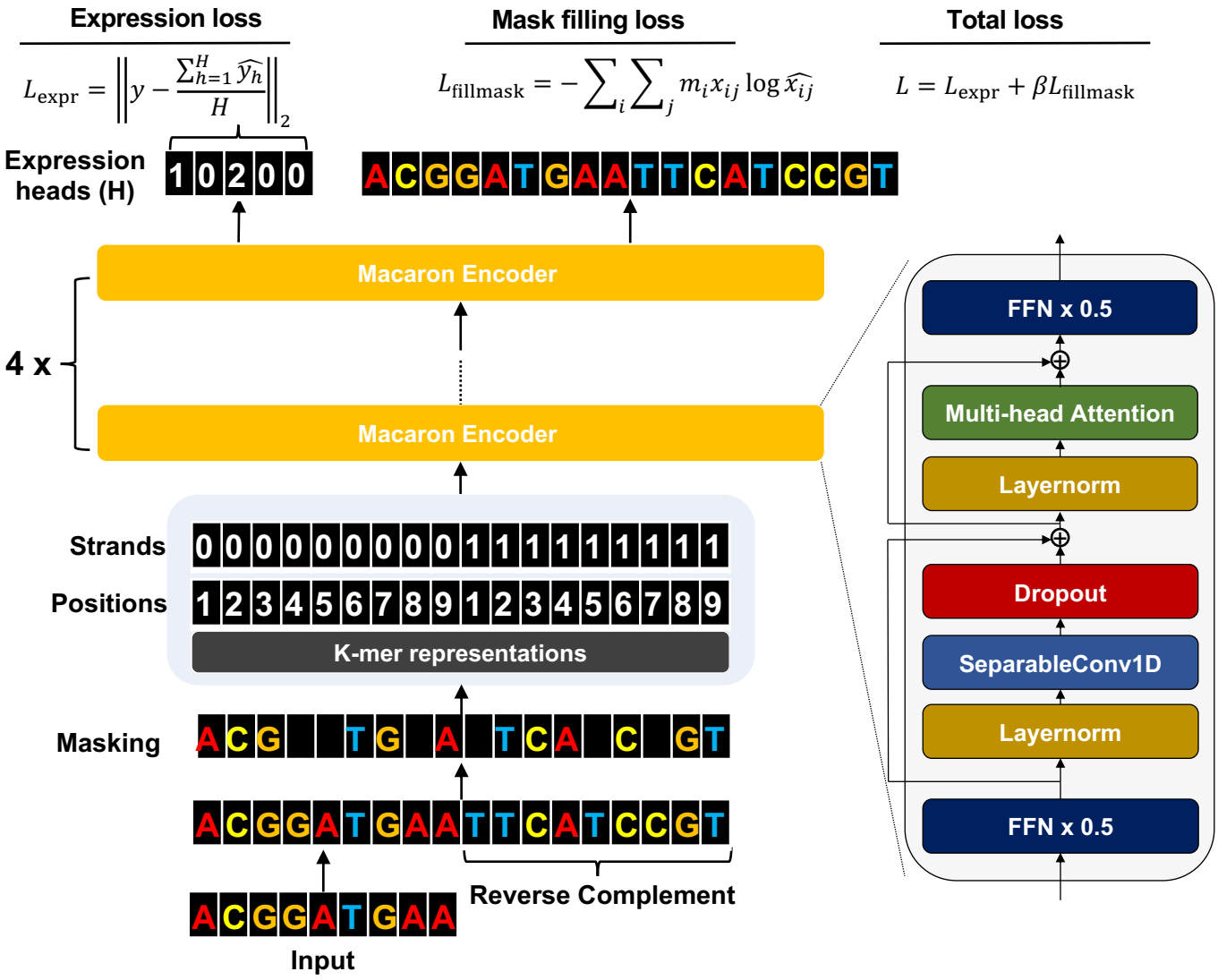
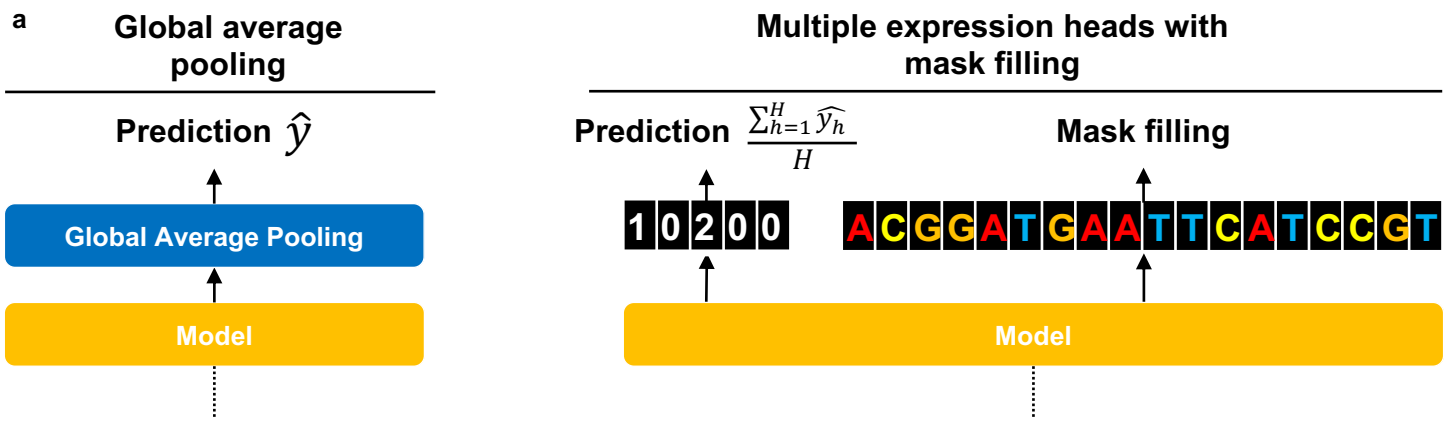
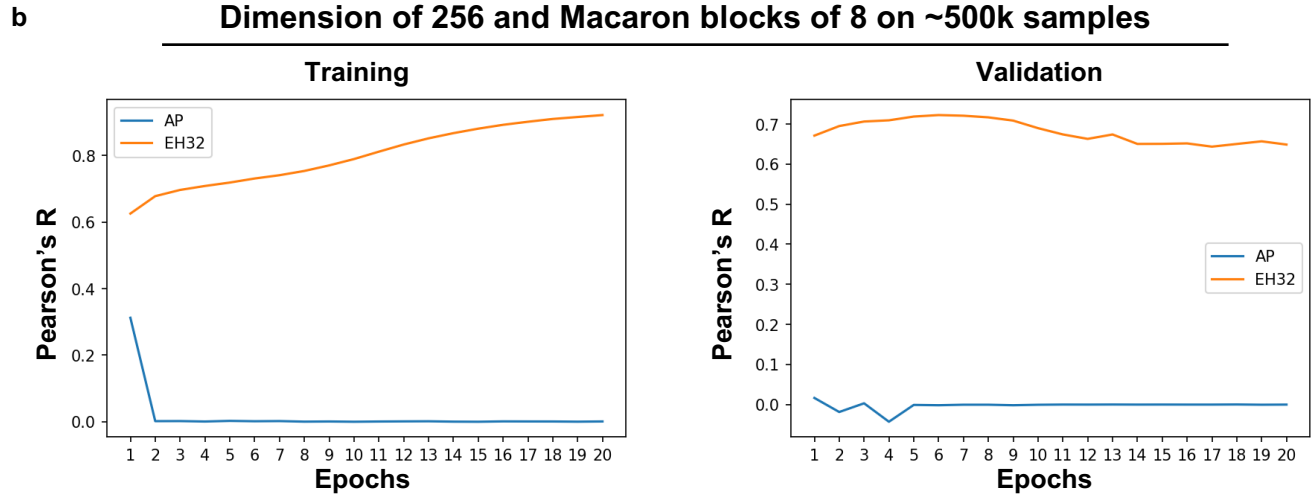


Figure 2



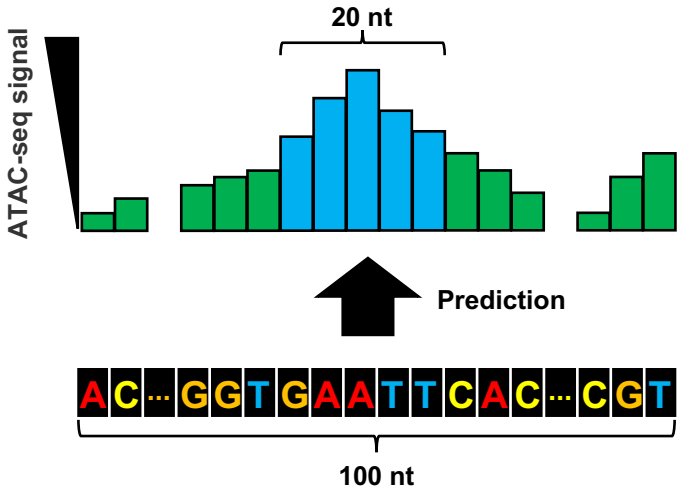


c

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

Figure 3

a

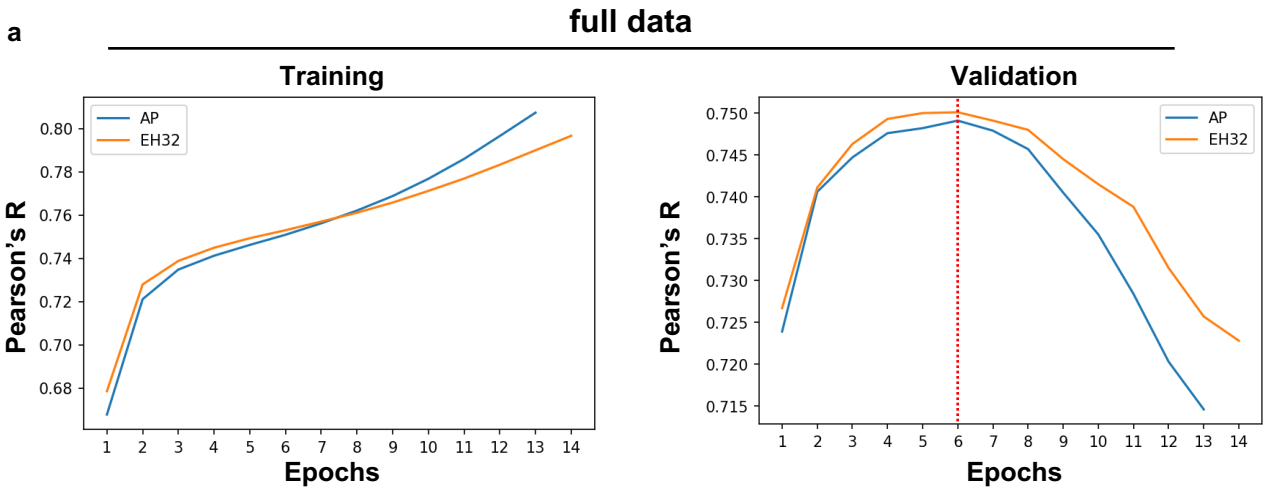


b

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

Figure 4

a



b

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 activation