MUSS: A Multi-scale and Sequence-based Model for Single-cell Gene Regulation

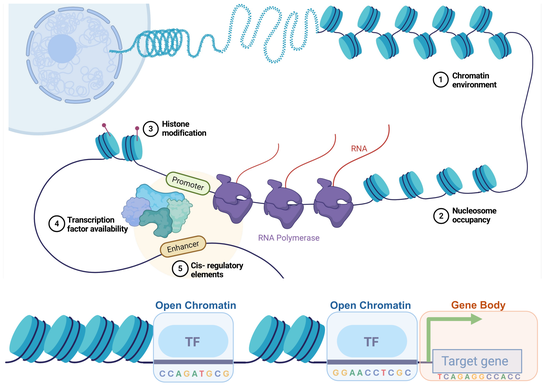
Kangjie Zheng[1*], Amirhossein Vahidi[1,2*], Liying Jin[1*], Joseph Clarke[1], Vijaya Baskar MS[1], Connor Rogerson[1], Mengji Zhang[1], Muzlifah Haniffa[1,3,4], Anshul Kundaje[6,7], Mo Lotfollahi[1,2,5,8†]

[1] Wellcome Sanger Institute [2] Cambridge Centre for AI in Medicine [3] Cambridge University Hospitals [4] CIITID, University of Cambridge
[5] Cambridge Stem Cell Institute [6] Department of Genetics, Stanford University [7] Department of Computer Science, Stanford University [8] Department of Medicine, University of Cambridge

Biological Motivation

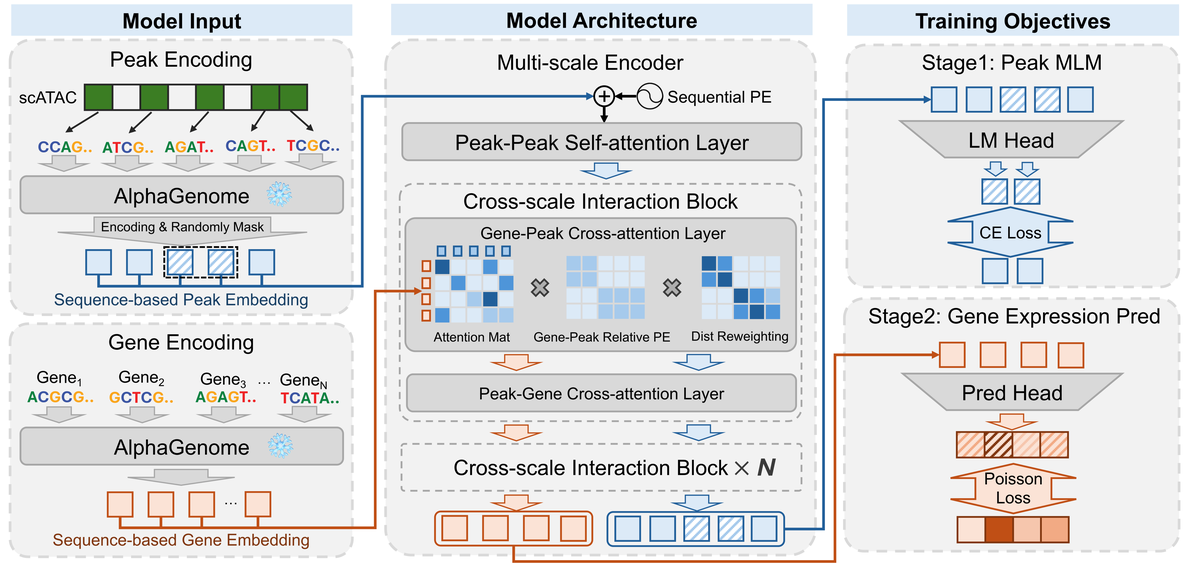
› Shared Genome, cell-specific expression

Cells share DNA sequence but differ in chromatin accessibility and gene expression. Transcription factors bind DNA motifs to regulate transcription.



How can we connect DNA sequence to cell-specific gene expression across regulatory scales?

Multi-scale Modeling and Training



Inputs: scATAC-seq (Stage 1); paired scATAC-seq + scRNA-seq (Stage 2)

› Shared DNA representations

A frozen AlphaGenome encoder produces both peaks and genes into a shared DNA-derived embedding space.

› Cell-specific multi-scale interactions

Peak self-attention and distance-biased, bidirectional peak-gene cross-attention.

› Two-stage training

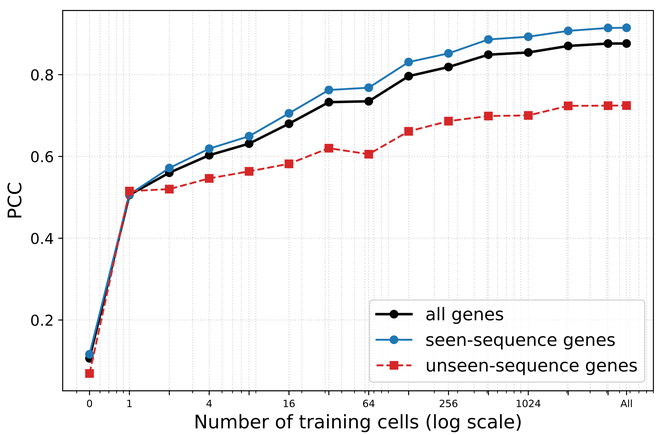
Peak MLM: Recover 15% masked peaks in PBMC169K scATAC data.

Expression prediction: Train on paired scATAC-scRNA data to predict gene expression with a Poisson loss.

Few-shot Adaptation

› Data-efficient expression prediction

One paired cell yields ~0.5 PCC after peak-MLM pretraining



PBMC10K; fine-tuning from the same peak-MLM checkpoint. Curves distinguish all, seen-sequence and unseen-sequence genes.

Experimental Results

› Expression beyond the training vocabulary

Unseen sequences: Highest R^2 in all 8 dataset x cell-type settings

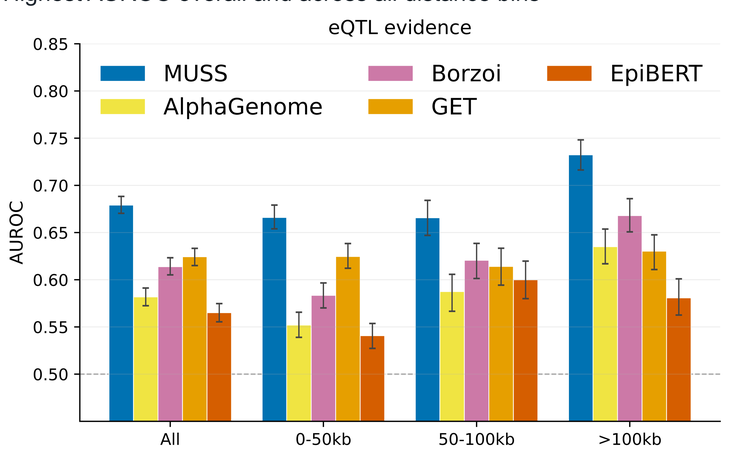
	Model	PBMC10K PCC↑	PBMC10K CCC↑	PBMC10K R^2↑	BMMC PCC↑	BMMC CCC↑	BMMC R^2↑	Openproblem PCC↑	Openproblem CCC↑	Openproblem R^2↑	Bone PCC↑	Bone CCC↑	Bone R^2↑
Unseen sequence													
Seen cell types.													
Seq	Borzoi	0.560	0.237	-0.293	0.539	0.424	0.025	0.613	0.486	0.242	0.487	0.313	-0.040
Seq	AlphaGenome	0.609	0.402	-0.005	0.650	0.450	0.144	**0.774**	0.690	0.522	0.615	0.545	0.262
Seq	scooby	0.533	0.353	-0.088	0.655	0.543	0.293	0.709	0.696	0.312	0.655	**0.648**	0.278
ATAC+Seq	GET	0.310	0.285	-0.094	0.216	0.180	-0.774	0.206	0.188	-0.208	0.203	0.103	-2.440
ATAC+Seq	EpiBERT	0.491	0.386	-0.063	0.549	0.540	0.098	0.595	0.571	0.310	0.475	0.448	0.040
	MUSS	**0.684**	**0.663**	**0.409**	**0.748**	**0.738**	**0.524**	0.755	**0.735**	**0.552**	**0.683**	0.611	**0.427**
Unseen cell types.													
Seq	Borzoi	0.571	0.303	-0.134	0.549	0.457	-0.011	0.606	0.469	0.227	0.483	0.315	-0.040
Seq	AlphaGenome	0.613	0.425	0.057	0.617	0.370	-0.052	**0.768**	0.676	0.506	0.591	0.519	0.221
Seq	scooby	0.439	0.319	-0.152	0.628	0.536	0.232	0.681	0.659	0.305	0.655	**0.648**	0.231
ATAC+Seq	GET	0.342	0.298	-0.187	0.196	0.168	-0.752	0.215	0.196	-0.208	0.201	0.110	-2.402
ATAC+Seq	EpiBERT	0.639	0.587	0.304	0.564	0.550	0.105	0.618	0.606	0.340	0.548	0.517	0.150
	MUSS	**0.698**	**0.686**	**0.436**	**0.742**	**0.724**	**0.481**	0.762	**0.742**	**0.564**	**0.683**	0.609	**0.419**

Code, datasets and model checkpoints

github.com/Lotfollahi-lab/MUSS_neurips26 🤗 huggingface.co/kangjiezheng/MUSS

› Enhancer-gene linking

Highest AUROC overall and across all distance bins



GeneHancer eQTL-supported links; balanced 1:1 chromosome- and distance-matched controls. Error bars: 95% bootstrap CIs.

* Equal contribution. † Correspondence: ml19@sanger.ac.uk. Biological illustration incorporates BioRender artwork.