DeepCNet: A multimodal deep learning model for predicting cell type-specific gene expression and promoter-enhancer interactions from single-cell multiome data

Weijia Jin[1], Fengdi Zhao[1], Xin Ma[1], Qing Lu[1], Karyn A. Esser[2,3], Ramon C. Sun[4,5,6,7], Li Chen[1,4,5,6]*

[1]Department of Biostatistics, University of Florida, Gainesville, FL, USA

[2]Department of Physiology and Aging, University of Florida, Gainesville, FL, USA

[3]Claude D. Pepper Older Americans Independence Center, University of Florida, Gainesville, FL

[4]Department of Biochemistry and Molecular Biology, University of Florida, Gainesville, FL, USA

[5]Center for Advanced Spatial Biomolecular Research, University of Florida, Gainesville, FL, USA

[6]Evelyn F. and William L. McKnight Brain Institute, University of Florida, Gainesville, FL, USA

[7]Cancer Institute, University of Florida, Gainesville, FL, USA

Correspondence: li.chen1@ufl.edu

Supplementary figures

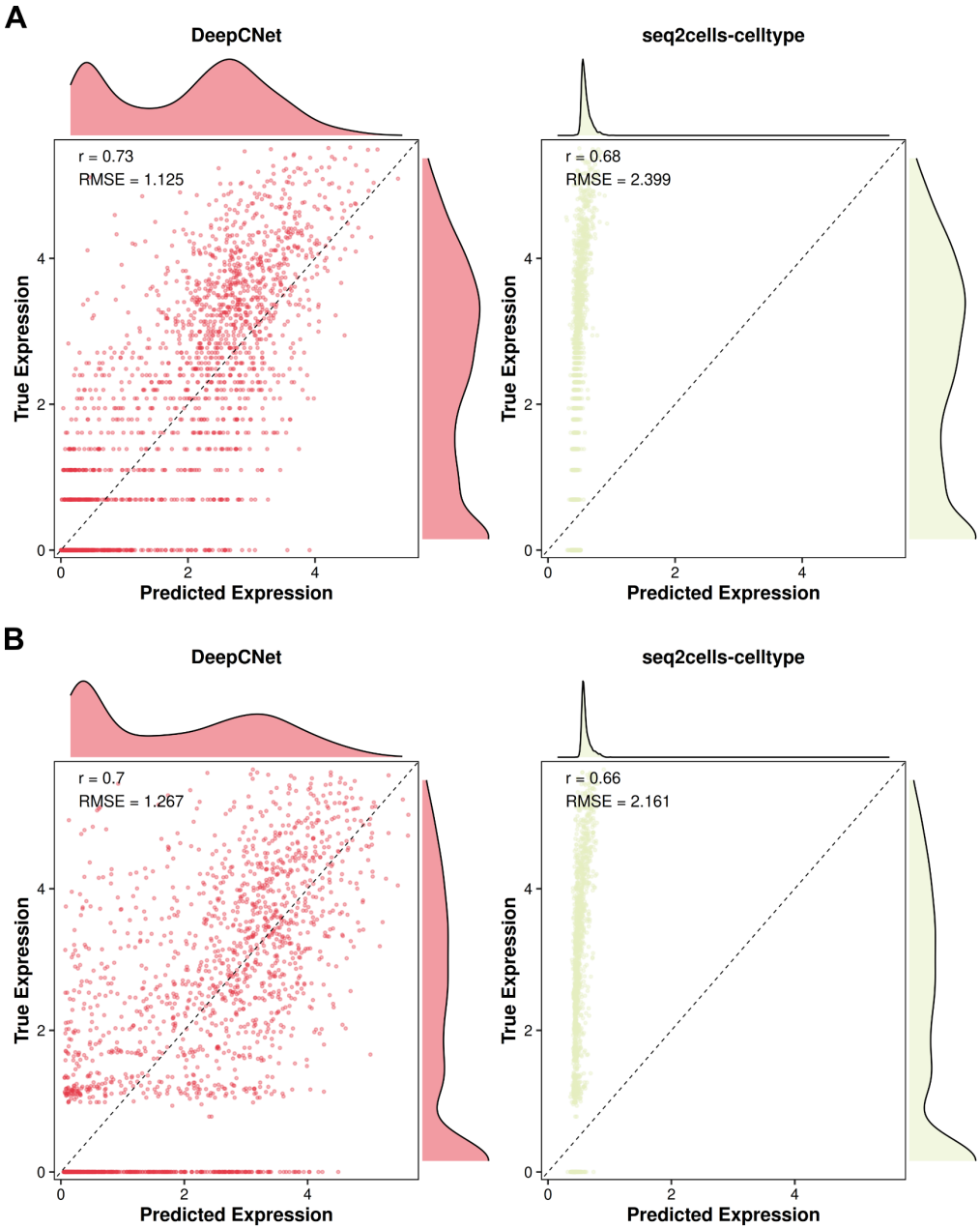

Figure S1. Representative examples of cell type-specific gene expression prediction. Scatter plots show observed versus predicted expression for held-out test genes from single-fold cross-chromosome validation. **A.** Regulatory T cells (Tregs) from the PBMC scMultiome dataset, comparing DeepCNet with seq2cells-celltype. **B.** Microglia from the MFC scMultiome dataset, comparing DeepCNet with seq2cells-celltype.

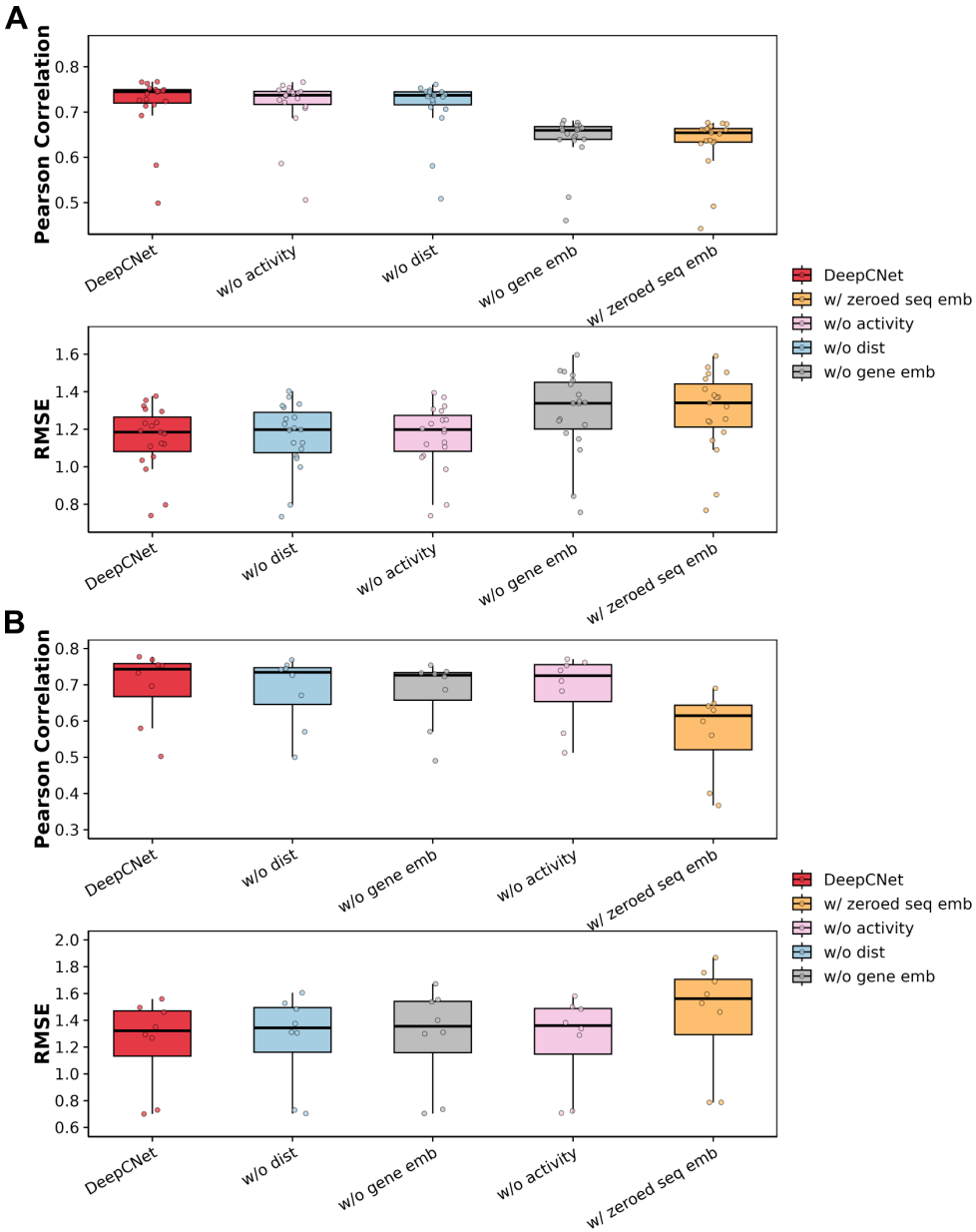

Figure S2. Ablation analyses of DeepCNet for gene expression prediction. DeepCNet full model was compared with four ablated variants: (i) w/o gene embedding, in which gene embeddings learnt by the cross-cell-type autoencoder were removed; (ii) w/o activity, in which ATAC-derived chromatin accessibility was removed; (iii) w/o distance, in which peak-to-TSS distance was removed; and (iv) w/ zeroed seq embedding, in which ExPecto-derived sequence embeddings were masked with zeros. Prediction performance was evaluated using 12-fold cross-chromosome validation strategy. Methods are ranked by decreasing median Pearson correlation or increasing median RMSE. **A.** PBMC scMultiome dataset. **B.** MFC scMultiome dataset.

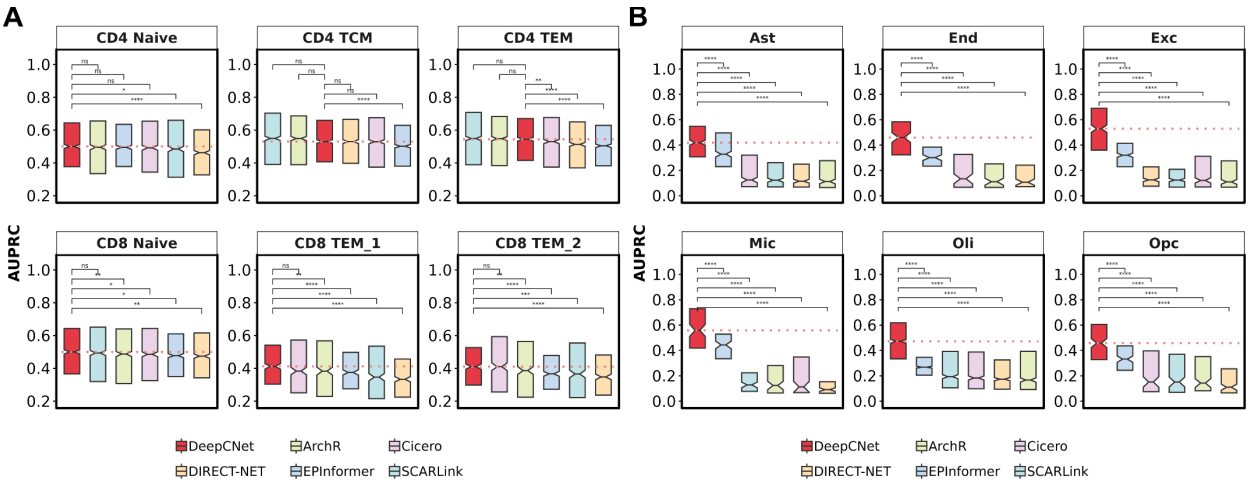

Figure S3. Benchmarking DeepCNet for predicting promoter–enhancer interactions by AUPRC. Inferred promoter–enhancer interactions overlapping experimentally validated chromatin interactions were treated as positives, while all remaining interactions were considered negatives. Analyses included only genes satisfying predefined inclusion criteria: more than 10 peaks; more than 5 positives and 5 negatives for the PBMC dataset; more than 3 positives and 3 negatives for the MFC dataset. For each gene, AUPRC was averaged across 12-fold cross-chromosome validation in the training set. Cell types are ranked by the median performance metric values. The AUROC/AUPRC across evaluated genes are visualized by interquartile range (IQR).**A.** AUPRC evaluation on the PBMC scMultiome dataset using cell type-matched ChIA-PET data. **B.** AUPRC evaluation on the MFC scMultiome dataset using cell type-matched sn-m3C-seq data.

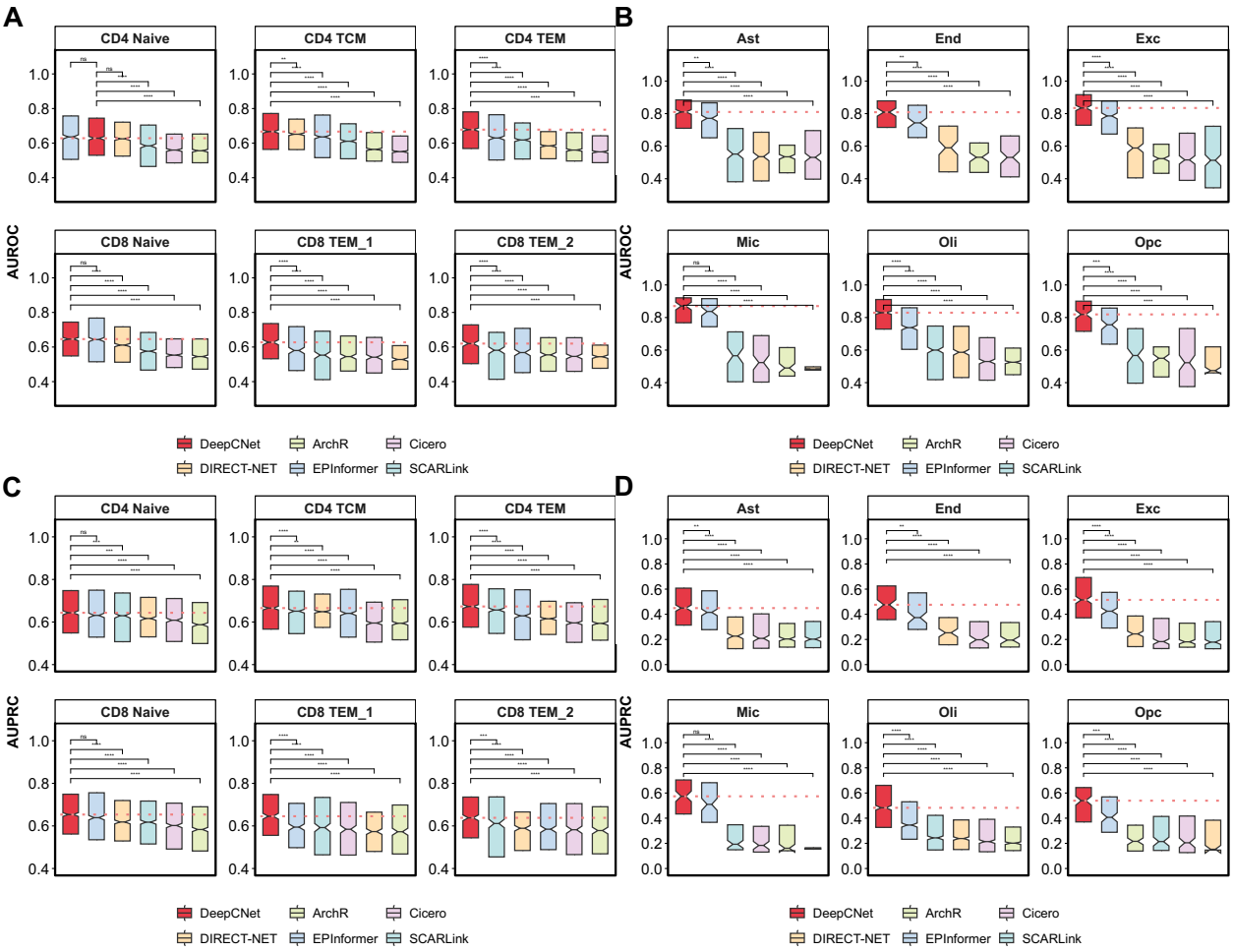

Figure S4. Benchmarking DeepCNet for promoter–enhancer interaction prediction using distance-matching. Inferred promoter-enhancer interactions are overlapping experimentally validated chromatin interaction are treated as positives, while all remaining ones are considered negatives. Within each cell type, positive and negative promoter–enhancer pairs were further matched by peak-to-TSS distance using nearest-neighbor propensity score without replacement, at a 1:1 ratio for PBMC and a 1:5 ratio for MFC scMultiome datasets. Analyses include only genes that satisfy predefined inclusion criteria: more than 10 peaks; more than 5 positives and 5 negatives for the PBMC dataset; more than 3 positives and 3 negatives for the MFC dataset. For each gene, AUROC or AUPRC was averaged across 12-fold cross-chromosome validation in the training set. Cell types are ordered by median performance. The AUROC/AUPRC across evaluated genes are visualized by interquartile range (IQR). **(A, C)** AUROC and AUPRC, respectively, for the PBMC scMultiome dataset using cell type-matched ChIA-PET data. **(B, D)** AUROC and AUPRC, respectively, for the MFC scMultiome dataset using cell type-matched sn-m3C-seq data.

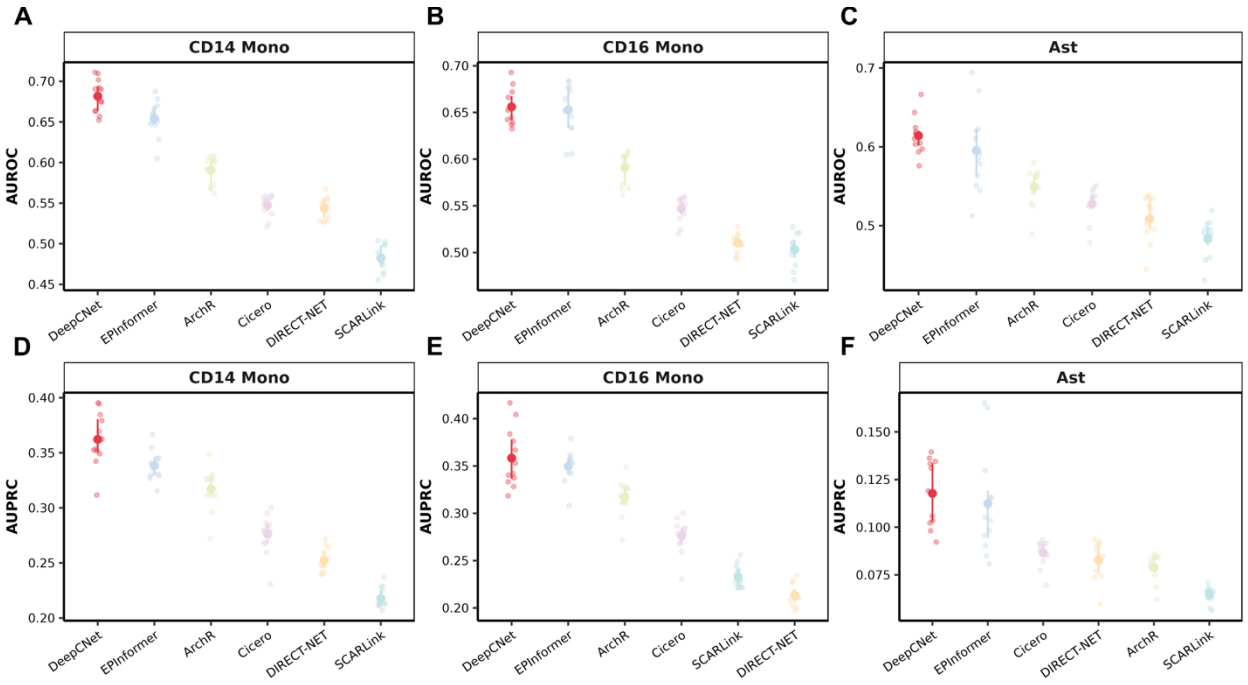

Figure S5. Benchmarking DeepCNet for predicting promoter-enhancer interaction using CRISPR perturbation experiments. Inferred promoter-enhancer interactions are labelled by overlapping with CRISPR-perturbed gene-enhancer pairs, allowing a 1kb positional gap. CD14 monocytes and CD16 monocytes from PBMC scMultiome dataset and Astrocyte from MFC scMultiome dataset are evaluated against CRISPR perturbation experiments performed in K562 and Astrocyte cell line respectively. Evaluation is performed using 12-fold cross-chromosome validation on the training genes. Vertical line indicates the interquartile range across folds. Methods are ordered by decreasing mean AUROC or AUPRC. **A.** AUROC evaluation on CD14 monocytes in PBMC scMultiome dataset. **B.** AUROC evaluation on CD16 monocytes in PBMC scMultiome dataset. **C.** AUROC evaluation on Astrocytes in MFC scMultiome dataset. **D.** AUPRC evaluation on CD14 monocytes in PBMC scMultiome dataset. E. AUPRC evaluation on CD16 monocytes in PBMC scMultiome dataset. F. AUPRC evaluation on Astrocytes in MFC scMultiome dataset.

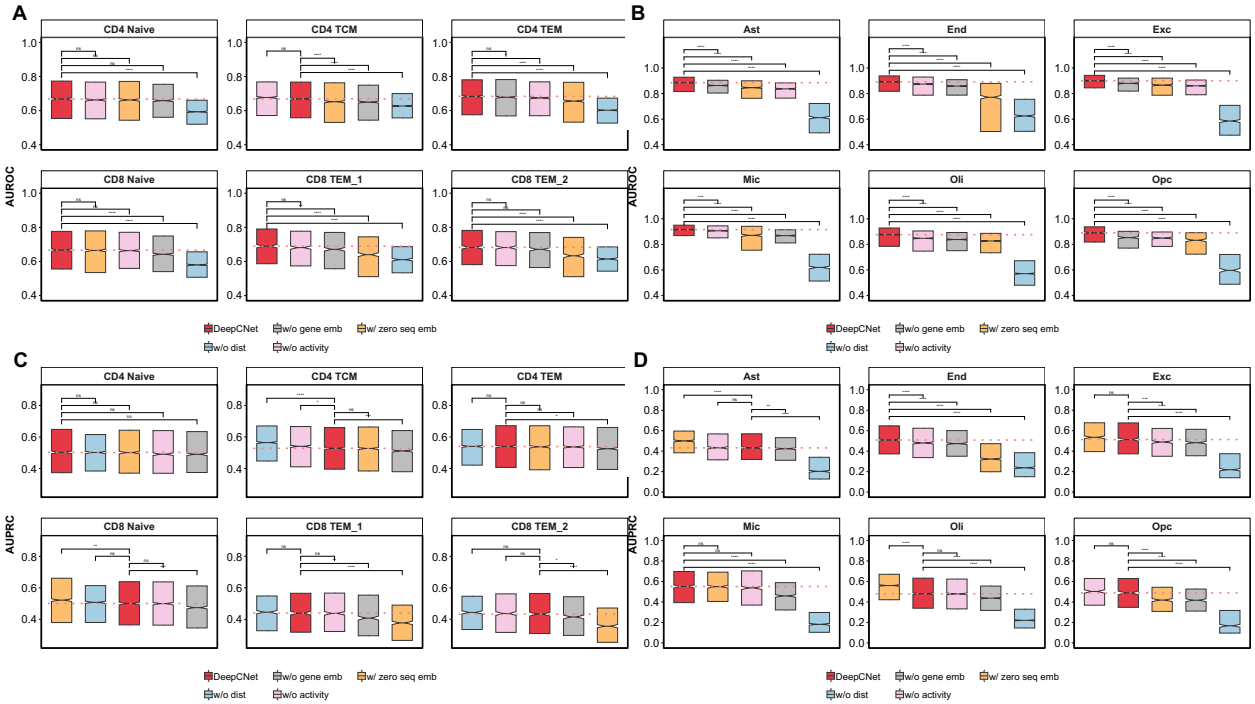

Figure S6. Ablation analysis of DeepCNet for promoter–enhancer interaction prediction. DeepCNet was compared with four ablated variants: w/o gene embedding, w/o activity, w/o distance, and w/ zeroed seq embedding. Inferred promoter–enhancer interactions overlapping experimentally supported chromatin interactions were treated as positives, while the remaining interactions were treated as negatives. For each gene, AUROC or AUPRC was averaged across cross-chromosome folds in the training set. Cell types are ranked by the median performance metric values. The AUROC/AUPRC across evaluated genes are visualized by interquartile range (IQR). **(A, B)** AUROC and AUPRC, respectively, for the PBMC scMultiome dataset using cell type-matched ChIA-PET data. **(C, D)** AUROC and AUPRC, respectively, for the MFC scMultiome dataset using cell type-matched sn-m3C-seq data.

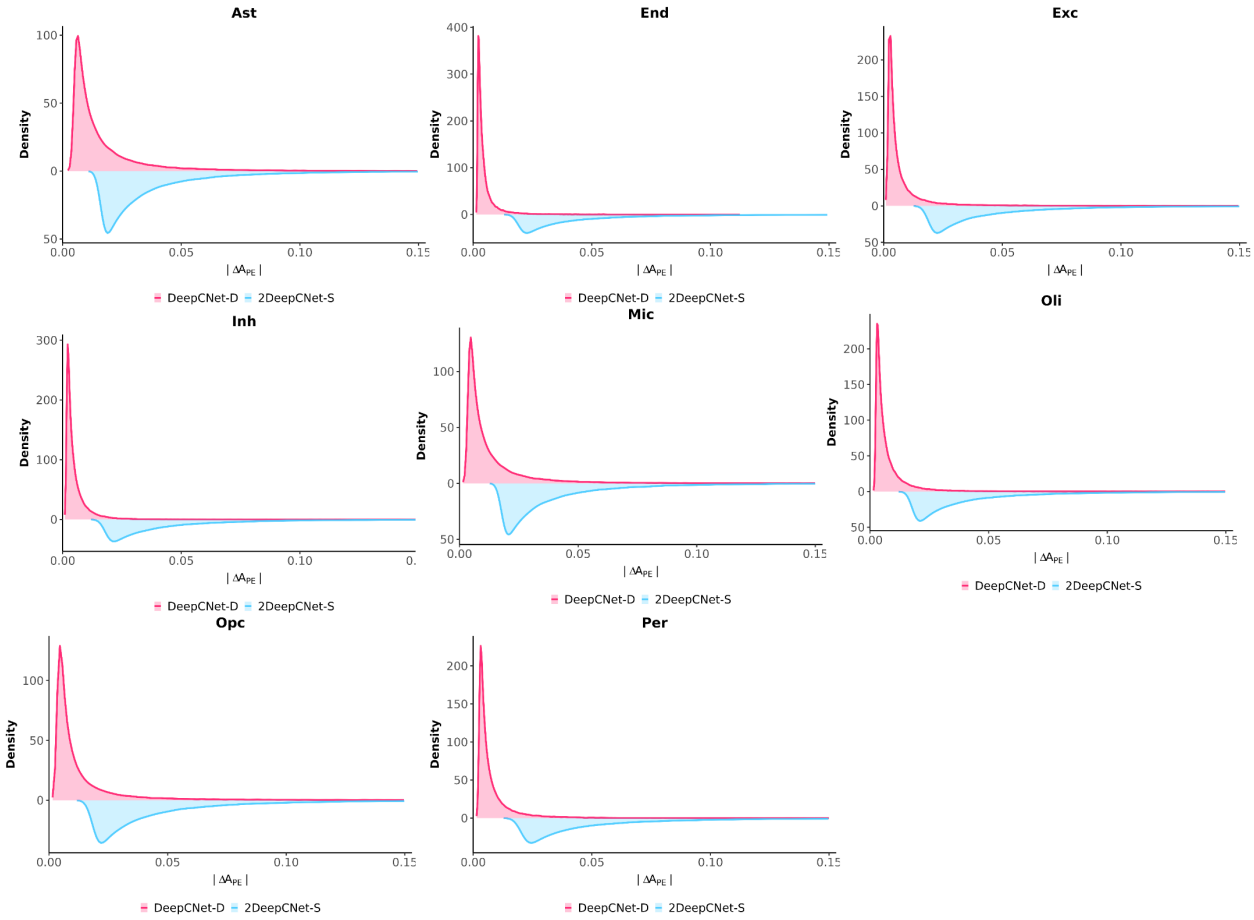

Figure S7. Density plots of the absolute values of DeepCNet-inferred differential promoter-enhancer attention scores (ΔA_{PE}) between AD and healthy control in eight cell types. Top 20% of absolute ΔA_{PE} values are shown. Red curves correspond to results obtained from a single joint model (DeepCNet-D), whereas blue curves represent results from modeling the two conditions separately (2DeepCNet-S).

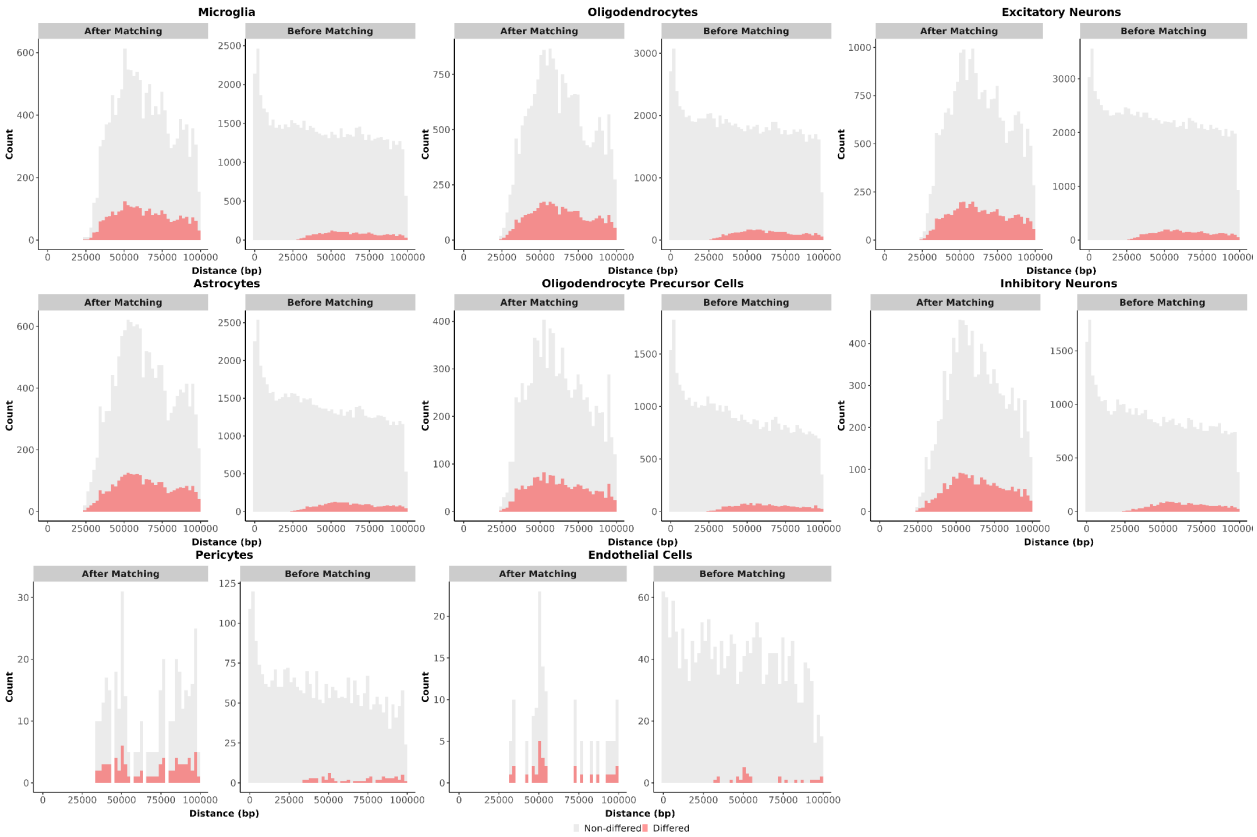

Figure S8. Density plots of TSS-to-peak distances before and after distance matching for differential and non-differential PE interactions across eight cell types in MFC scMutiome data. A PE interaction is defined as differential (Differed, red) if it is overlapped with Hi-C loops only in one condition, allowing a 1 kb gap, and as non-differential (Non-differed, gray) if overlapped with neither Hi-C conditions or overlapped with both conditions.

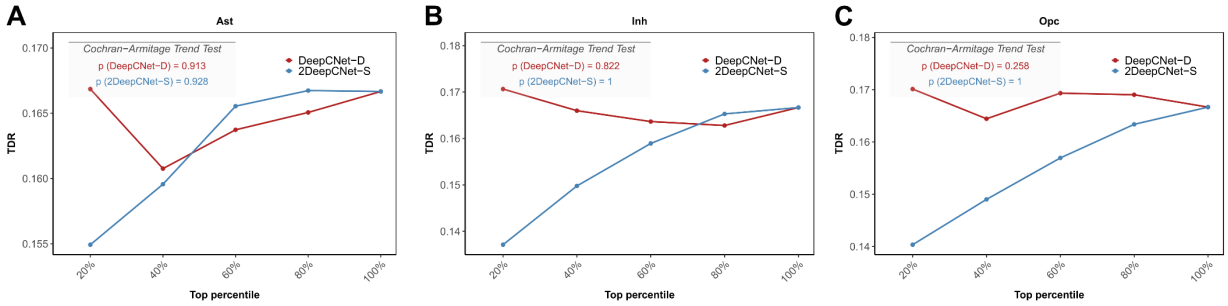

Figure S9. Enrichment plots showing the true discovery rate (TDR) of top inferred differential PE interactions validated by two-conditional Hi-C loops in three cell types (Ast, Inh, and Opc). P-values are calculated using a one-sided Cochran-Armitage trend test for a decreasing trend across five percentiles (20th, 40th, 60th, 80th, and 100th).

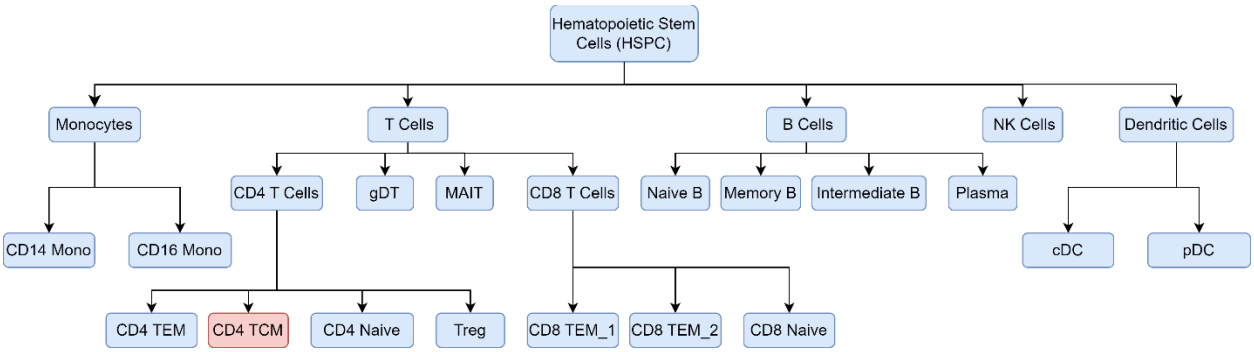

Figure S10. Cell type hierarchy of the PBMC scMultiome dataset. The cell type highlighted in red represents the anchor cell type.

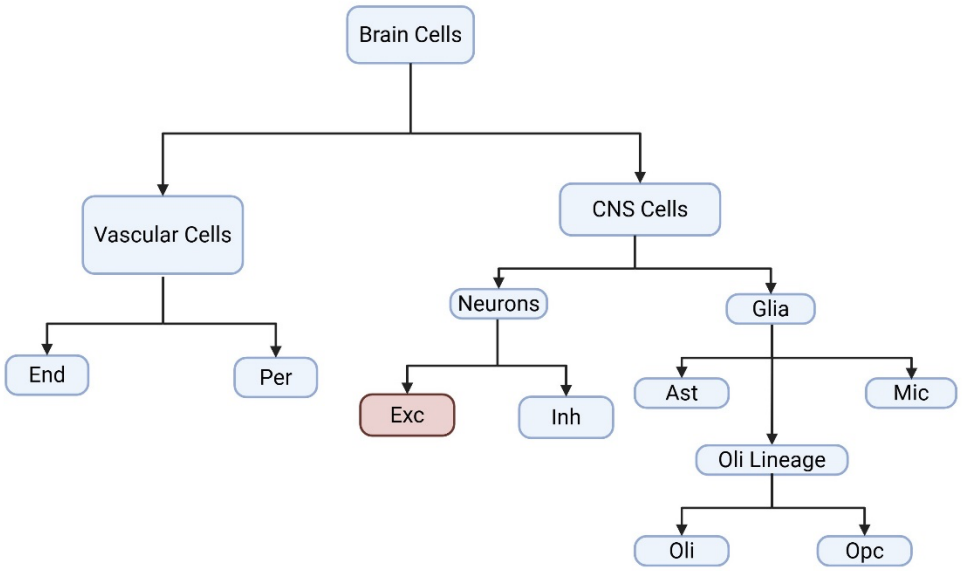

Figure S11. Cell type hierarchy of the MFC scMultiome dataset. The cell type highlighted in red represents the anchor cell type.

Supplementary tables

Cell Type	Full Name	Sample Size (AD_Normal)	Sample Size (AD_Disease)
Ast	Astrocytes	584	248
End	Endothelial cells	14	4
Exc	Excitatory neuron	2362	1203
Inh	Inhibitory neuron	1369	635
Mic	Microglia	277	237
Oli	Oligodendrocyte	1509	2649
Opc	Oligodendrocyte progenitor cells	331	163
Per	Pericyte	18	14

Table S1. Cell type composition and sample size of MFC scMultiome dataset.

Cell Type	Sample Size
HSPC	23
CD14 Mono	2790
CD16 Mono	521
CD4 Naïve	1332
CD4 TCM	1180
CD4 TEM	542
Treg	182
CD8 Naïve	1466
CD8 TEM_1	302
CD8 TEM_2	363
MAIT	108
gdT	107
Naive B	145
Intermediate B	357
Memory B	368
Plasma	18
pDC	85
cDC	198
NK	445

Table S2. Cell type composition and sample size of the PBMC scMultiome dataset.

Cell Type	ENCODE File
CD8 TEM_1	ENCFF004MDS
CD8 TEM_2	ENCFF004MDS
CD4 TCM	ENCFF877IHO
CD4 TEM	ENCFF877IHO
CD4 Naive	ENCFF877ABS
CD8 Naive	ENCFF218BSY

Table S3. Six cell types from PBMC scMultiome dataset and their matched ENCODE ChIA-PET data.

Cell Type	Full Name	sn-m3C-seq Contact File
Ast	Astrocytes	Astro_all_brain.txt_1kb_contacts.mcool
End	Endothelial cells	Endo_all_brain.txt_1kb_contacts.mcool
Exc	Excitatory neuron	L2-3_all_brain.txt_1kb_contacts.mcool
Mic	Microglia	MG_all_brain.txt_1kb_contacts.mcool
Oli	Oligodendrocyte	ODC_all_brain.txt_1kb_contacts.mcool
Opc	Oligodendrocyte progenitor cells	OPC_all_brain.txt_1kb_contacts.mcool

Table S4. Six cell types from MFC_Normal scMultiome dataset and their matched sn-m3C-seq chromatin contact data.