Supplemental Materials for

Biologically faithful bidirectional translation between single-cell transcriptomes and DNA methylomes with scBOND

Kehan Lang[1,†], Chenyang Jia[1,†], Siyu Li[1,†], Yi Guo[2], Dingjun Hu[1] and Shengquan Chen[1,3,*]

[1] School of Mathematical Sciences and LPMC, Nankai University, Tianjin 300071, China

[2] College of Electronic Information and Optical Engineering, Nankai University, Tianjin 300350, China

[3] Academy for Advanced Interdisciplinary Studies, Nankai University, Tianjin 300071, China

[†] These authors contributed equally: Kehan Lang, Chenyang Jia and Siyu Li

* Correspondence to: Shengquan Chen **(Email:** chenshengquan@nankai.edu.cn)

Index

Supplemental Texts

Text S1: Details of evaluation metrics

To quantitatively evaluate how well translated profiles preserve cellular heterogeneity, we assessed clustering results using four metrics: Adjusted Mutual Information (AMI), Adjusted Rand Index (ARI), Homogeneity (HOM), and Normalized Mutual Information (NMI) (Tu et al. 2022). AMI quantifies the agreement between predicted clusters and ground-truth cell-type labels by measuring shared information between the two labelings, while correcting for chance agreements. This adjustment makes AMI particularly suitable for datasets with imbalanced cell type distributions. ARI evaluates the similarity between cluster assignments by penalizing random overlaps and rewarding concordant groupings, particularly effective for datasets with balanced cell-type proportions. HOM assesses the purity of clusters by measuring the extent to which each predicted cluster contains cells from a single true cell type, with higher values indicating better preservation of biological heterogeneity. NMI further normalizes mutual information, providing an interpretable scale for comparing clustering quality across datasets. These metrics collectively provide a comprehensive assessment of translation accuracy, capturing both global alignment and fine-grained cellular distinctions.

Consider a set of translated profiles containing N cells with true cell type labels $T = \{T_1, \dots, T_n\}$ and cluster labels $P = \{P_1, \dots, P_m\}$ generated by the Leiden algorithm (Traag

et al. 2019). Let a_i be the number of cells belonging to cell type T_i, b_j the number of cells in cluster P_j, and n_{ij} the number of cells common to both T_i and P_j. AMI further accounts for chance agreement in mutual information:

$$AMI = \frac{MI(P,T) - E[MI(P,T)]}{\frac{1}{2}\big(H(P) + H(T)\big) - E[MI(P,T)]}.$$

where $MI(\cdot)$ denotes mutual information and $H(\cdot)$ represents entropy. ARI adjusts the Rand Index (RI) for chance agreement:

$$ARI = \frac{\sum_{i,j} n_{ij} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/(N)}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right]/(N)},$$

HOM measures the purity of cell types within clusters:

$$HOM = 1 - \frac{H(T|P)}{H(T)},$$

where $H(X|Y)$ is the conditional entropy of X given Y. NMI is a normalized form of MI:

$$NMI = \frac{MI(P,T)}{\sqrt{H(P)H(T)}}.$$

To further explore transcriptional-epigenetic interactions, we computed Spearman's correlation coefficients between gene expression levels and promoter-proximal DNA methylation levels for individual cells. Spearman's correlation coefficients evaluate the performance of translation from a numerical accuracy standpoint, given by the following

formulas:

$$Spearman\ index\ =\ \frac{1}{N}\sum_{i=1}^{N} Spearman\ (\mathbf{X}_{\mathrm{r}}[i], \mathbf{X}_{\mathrm{r}\to\mathrm{m}}^{\mathrm{pred}}[i]),$$

where $\mathbf{X}_{\mathrm{r}}[i]$ and $\mathbf{X}_{\mathrm{r}\to\mathrm{m}}^{\mathrm{pred}}[i]$ represent the gene expressions and predicted promoter-proximal DNA methylation levels of cell i, respectively, and N represents the total count of cells.

Text S2: Implementation details of downstream analyses

We identified the differentially expressed genes (DEGs) using the SCANPY pipeline (Wolf et al. 2018) and differentially methylated regions (DMRs) using the EpiScanpy pipeline (Danese et al. 2021).

Statistical test methods for identifying differentially expressed genes and methylated regions

For both DEGs and DMRs, we used the sc.tl.rank_genes_groups function in SCANPY [https://github.com/scverse/scanpy/blob/1.11.5/src/scanpy/tools/_rank_genes_groups.py] to perform ranking and statistical testing of genes across conditions. This function is a robust tool for differential analysis in scRNA-seq and other omics data.

For differentially expressed genes (DEGs), we applied the Wilcoxon rank-sum test. This test is chosen for its non-parametric nature, which makes it particularly suitable for scRNA-seq data, which is often characterized by sparse, discrete counts and overdispersion. The Wilcoxon test compares the ranks of expression values between two groups and is robust to the non-normality and heterogeneity of scRNA-seq data, making it well-suited for identifying genes that are differentially expressed in a non-parametric fashion.

For differentially methylated regions (DMRs), we used the t-test. The t-test was selected due to the continuous nature of DNA methylation data, where methylation levels are measured on a continuous scale between 0 and 1. The t-test is parametric and assumes

that the data follows a normal distribution, which is typically valid for methylation data after transformation (such as logit transformation) or when the sample size is large enough to invoke the central limit theorem. This allows for a direct comparison of methylation levels between groups, providing statistical rigor for identifying significant DMRs.

Enrichment analysis of DEGs

We used the DEGs to perform gene enrichment analysis using the DAVID tool (Huang et al. 2007), covering both Gene Ontology (GO) (Consortium 2004) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa and Goto 2000) enrichment analyses. The GO terms include molecular function, biological process, and cellular component.

Downstream analysis of DMRs

LDSC: Partitioned linkage disequilibrium score regression (LDSC) is a statistical method used to quantify the contribution of genetic variants to the heritability of traits or diseases across the genome (Finucane et al. 2015). Developed to interpret genome-wide association study (GWAS) results, LDSC leverages linkage disequilibrium information, the non-random association of alleles at different loci, to differentiate the signal due to polygenic traits from confounding biases such as population stratification and cryptic relatedness.

SNPsea: SNPsea is an enrichment algorithm designed for analyzing single-nucleotide polymorphisms (SNPs) to pinpoint specific cell types, tissues, and biological pathways that are influenced by risk loci associated with traits (Slowikowski et al. 2014). It tests trait-

associated genomic loci for enrichment of specificity to conditions (cell types, tissues and pathways). We quantified the enrichments of cell-type-specific peaks in tissue-specific accessibility profiles across 79 tissues. The top 30 significantly enriched tissues are illustrated.

Text S3: Details of sequencing experiment error mimicking strategies

Details of value flipping methods for DNA methylation error simulation

Value flipping was implemented by transforming randomly selected methylation values using the operation $\mathbf{X}^{\mathrm{m}}_{\mathrm{reversed}}[i][j] = 1 - \mathbf{X}^{\mathrm{m}}_{\mathrm{processed}}[i][j]$, in which (i, j) refers to the random site for error simulation and $\mathbf{X}^{\mathrm{m}}_{\mathrm{processed}}$ refers to the DNA methylation expression matrix after data preprocessing. This approach consistently inverts methylation levels regardless of whether the data is represented as binary states (0/1) or continuous values (0-1 range).

Details of value dropout methods for RNA error simulation

Dropout strategy was executed as putting randomly selected RNA values $\mathbf{X}^{\mathrm{r}}_{\mathrm{reversed}}[i][j]$ to zero, thus mimicking dropout events common in scRNA-seq.

Text S4: Details of scBOND-Aug downstream cell type identification task

To evaluate the cell-type prediction accuracy of translated profiles, a downstream analysis was performed using the HumanBrainA dataset, which contains expert-annotated cell types with matched scRNA-seq and scDNAm profiles. For the RNA-based annotation task, both modalities were split into training and test sets at a 4:1 ratio. Random Forest (RF) and Support Vector Classifier (SVC) (Abdelaal et al. 2019) were trained on the original RNA training data to predict cell type labels. During testing, the DNAm test profiles were translated into RNA space using the baseline scBOND model and the augmented scBOND-Aug model. These translated profiles were then classified by the pre-trained RF and SVC models. Performance was assessed using four metrics: Accuracy, Cohen's Kappa, F1-Macro, and F1-Weighted.

As shown in Supplemental Fig. S10, scBOND-Aug consistently achieved stronger performance across all evaluation metrics and with both classifiers when compared to the baseline scBOND model. Improvements were observed in both translation directions, with particularly notable gains in the RNA-to-DNAm direction when using the SVC classifier. A similar consistent trend was evident with the RF classifier, indicating robust enhancement in prediction fidelity. These results suggest that translations produced by scBOND-Aug better preserve biologically discriminative features, thereby supporting more accurate cell-type identification in settings where only a single modality is available.

Text S5: Details of the robustness test of scBOND-Aug under cell-type perturbation

The label shuffling augmentation method is designed to evaluate model robustness by randomly perturbing cell type annotations while preserving the overall cellular composition. Specifically, given a shuffle rate parameter (ranging from 1% to 5% in our experiments), the method randomly selects a proportion of cells from the training set and reassigns their cell type labels to different categories, ensuring that the new labels are distinct from the original ones. To maintain the global cell type distribution, the algorithm employs a balanced reassignment strategy: it first removes the original labels of shuffled cells from the available pool, then redistributes alternative labels while guaranteeing that each reassigned label differs from the original. This approach creates controlled label noise that mimics potential annotation errors or biological heterogeneity, allowing for rigorous assessment of model performance under imperfect ground truth conditions while maintaining the biological relevance of the cellular composition.

Notably, scBOND-Aug demonstrated consistent performance stability across varying noise levels and surpassed the original scBOND model (which does not employ augmentation) in both translation directions (Supplementary Fig. S11). This indicates that the model exhibits robustness to moderate label mis-annotation, a property likely attributable to the augmentation strategy being confined to within-cell-type groups.

Consequently, random label flips affect only a minor fraction of the synthetically generated pairs.

Text S6: Details on the matrix dimensions of the self-attention mechanism

The attention mechanism in scBOND uses 8-head attention. For an input tensor $\mathbf{X} \in \mathbb{R}^{\text{batchsize}\times 128}$, we first introduce a sequence dimension, transforming the input to $\mathbf{X}' \in \mathbb{R}^{\text{batchsize}\times 1\times 128}$. The query ($Q$), key ($K$), and value ($V$) matrices are then projected into 8 attention heads, with each head having a dimension of 16 (since 128 / 8 = 16). This results in an attention output tensor $\mathbf{A} \in \mathbb{R}^{\text{batchsize}\times 8\times 1\times 128}$. The final output tensor is then integrated with the original latent representations to enhance the model's ability to capture complex relationships across modalities.

Text S7: Details of sampling methods

Interval partitioning: The continuous range [0.0, 1.0] is partitioned into four distinct intervals.

Label type assignment: Each interval is assigned a specific label type. The intervals [0.0, 0.2] and (0.8, 1.0] are designated for soft labels (retaining the original sampled random value), simulating positive and negative examples with higher uncertainty. The intervals (0.2, 0.5] and (0.5, 0.8] are designated for hard labels, fixed at 0.2 and 0.8, respectively, simulating high-confidence negative and positive examples.

Probability weighting: Each interval is assigned a specific probability weight for being sampled: 20% (soft negative), 30% (hard negative), 30% (hard positive), and 20% (soft positive).

Sampling execution: For each label to be generated, an interval is first randomly selected based on the predefined probability weights. A value is then uniformly sampled from within the selected interval. If the interval is designated for hard labels, this sampled value is replaced by the fixed value (0.2 or 0.8).

The key advantage of this strategy is that it allows for precise control over the proportion of hard vs. soft labels and positive vs. negative examples in the training data. By exposing the model to a mixture of samples with clear supervisory signals (hard labels)

and those with inherent uncertainty (soft labels), we facilitate the learning of more robust and generalizable feature representations.

Text S8: Details of constructing scDNAm data matrix

To construct the cell-by-region scDNAm data matrix, we begin by applying a sliding window, denoted as W, over the DNA sequence. For each window, we first count the number of CG sites within that window, represented as CG^{W}. Additionally, we compute two key quantities: the total number of observations within the window, denoted as $\mathrm{Tot}^{\mathrm{W}}$, and the number of observations in which DNA methylation is detected at the CG sites, represented as $\mathrm{Met}^{\mathrm{W}}$. If the number of CG sites CG^{W} in a window is smaller than a predefined threshold (set to 0), the DNA methylation value for that window is assigned as "NA" (not available). If the number of CG sites meets or exceeds the predefined threshold, the methylation ratio for the window is calculated as $\mathrm{Met}^{\mathrm{W}}/\mathrm{Tot}^{\mathrm{W}}$. This process is repeated for each cell, generating the final scDNAm data matrix by calculating statistics for each sliding window within each cell.

Supplemental Figures

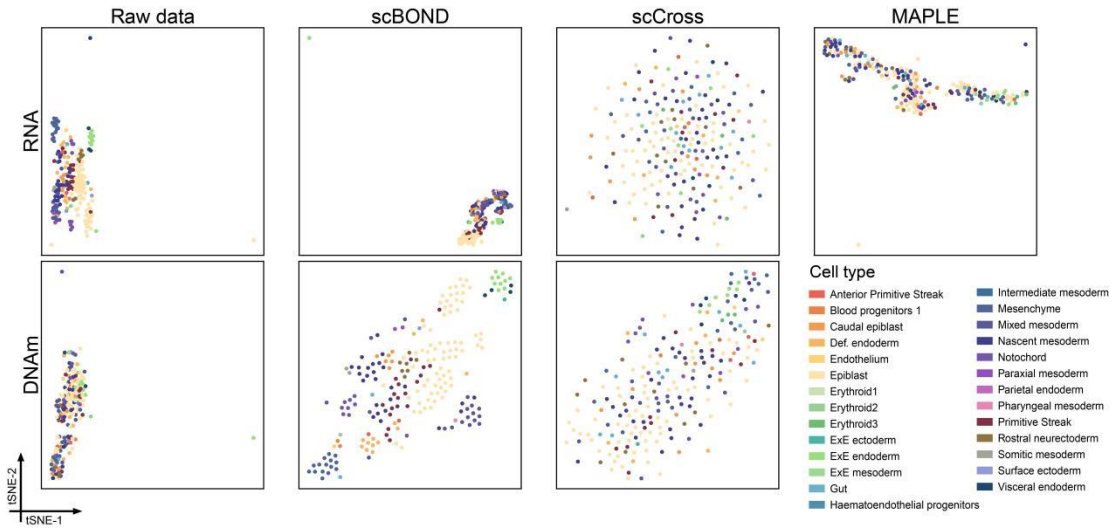

Figure S1. tSNE visualizations of the translated data from the first test fold of the MouseEmbryo dataset.

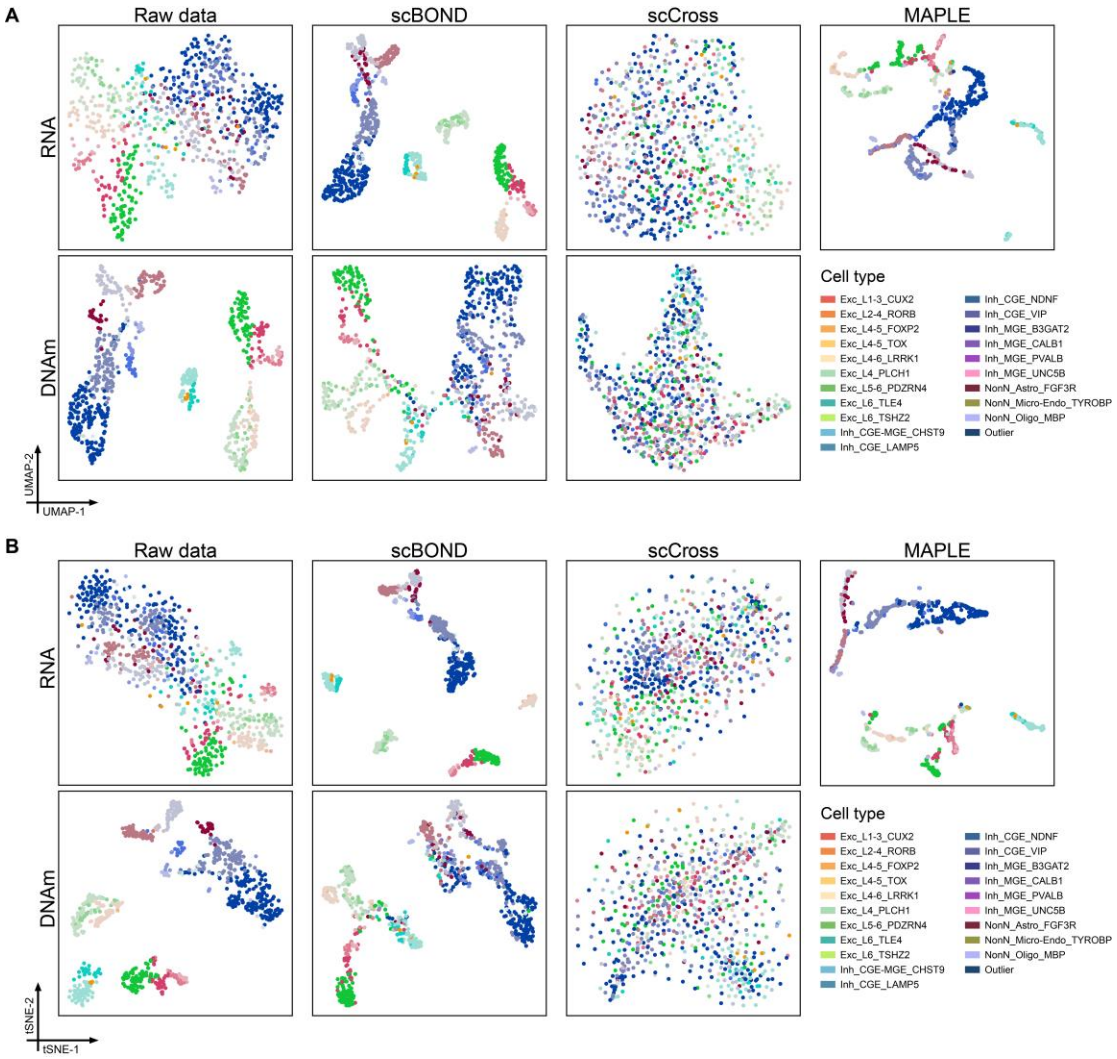

Figure S2. Visualizations on the HumanBrainA dataset. A, UMAP visualizations of the translated data from the first test fold on the HumanBrainA dataset. **B**, tSNE visualizations of the translated data from the first test fold on the HumanBrainA dataset.

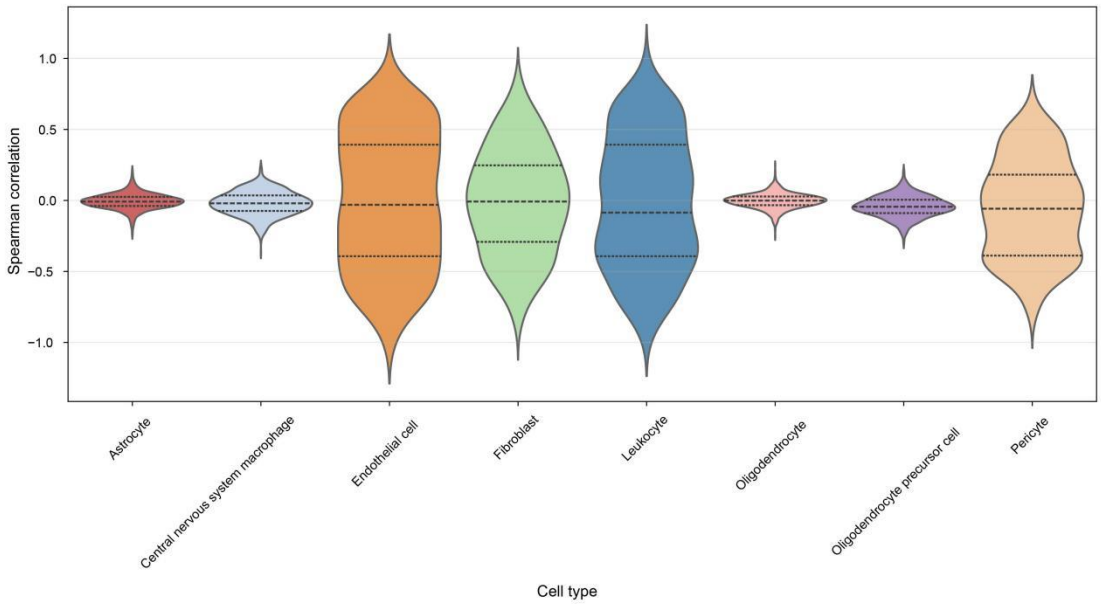

Figure S3. Distribution of Spearman's correlation coefficients between gene expression and promoter-proximal DNA methylation levels for different cell types on the HumanBrainB dataset.

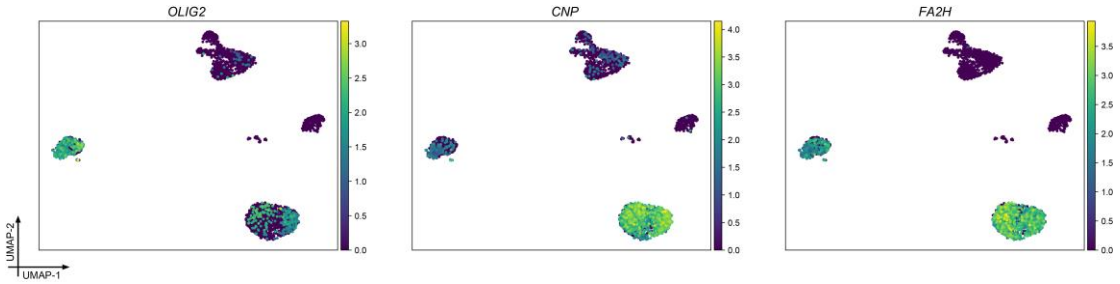

Figure S4. UMAP visualizations of gene expression for *OLIG2*, *CNP*, and *FA2H* on the HumanBrainB-RNA dataset.

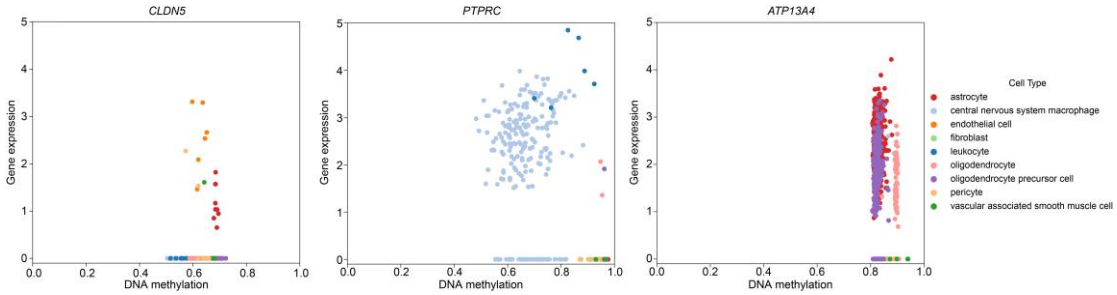

Figure S5. Scatter plots showing the relationship between gene expression and promoter-proximal DNA methylation levels across cell types for three representative marker genes of other major cell types: *CLDN5*, *PTPRC*, and *ATP13A4* on the HumanBrainB dataset.

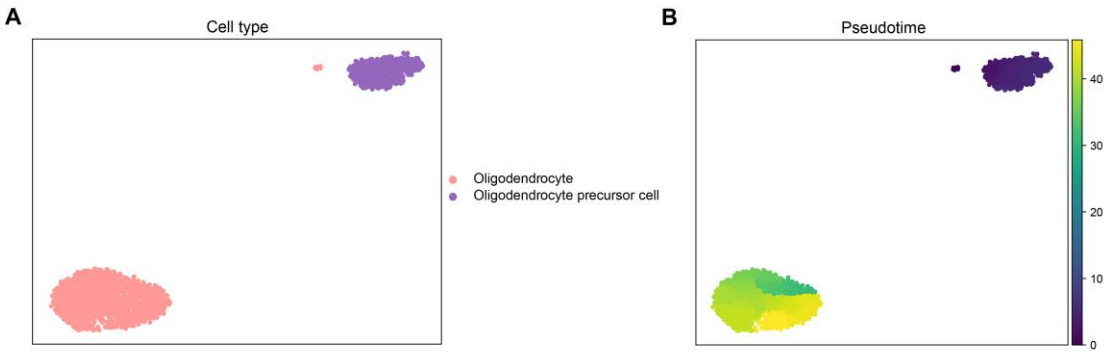

Figure S6. UMAP visualizations of oligodendrocytes and oligodendrocyte precursor cells in the HumanBrainB-RNA dataset and pseudotime trajectory analysis. A, UMAP visualizations of oligodendrocytes and oligodendrocyte precursor cells on the HumanBrainB-RNA dataset. **B,** UMAP visualizations of pseudotime trajectory on oligodendrocytes and oligodendrocyte precursor cells.

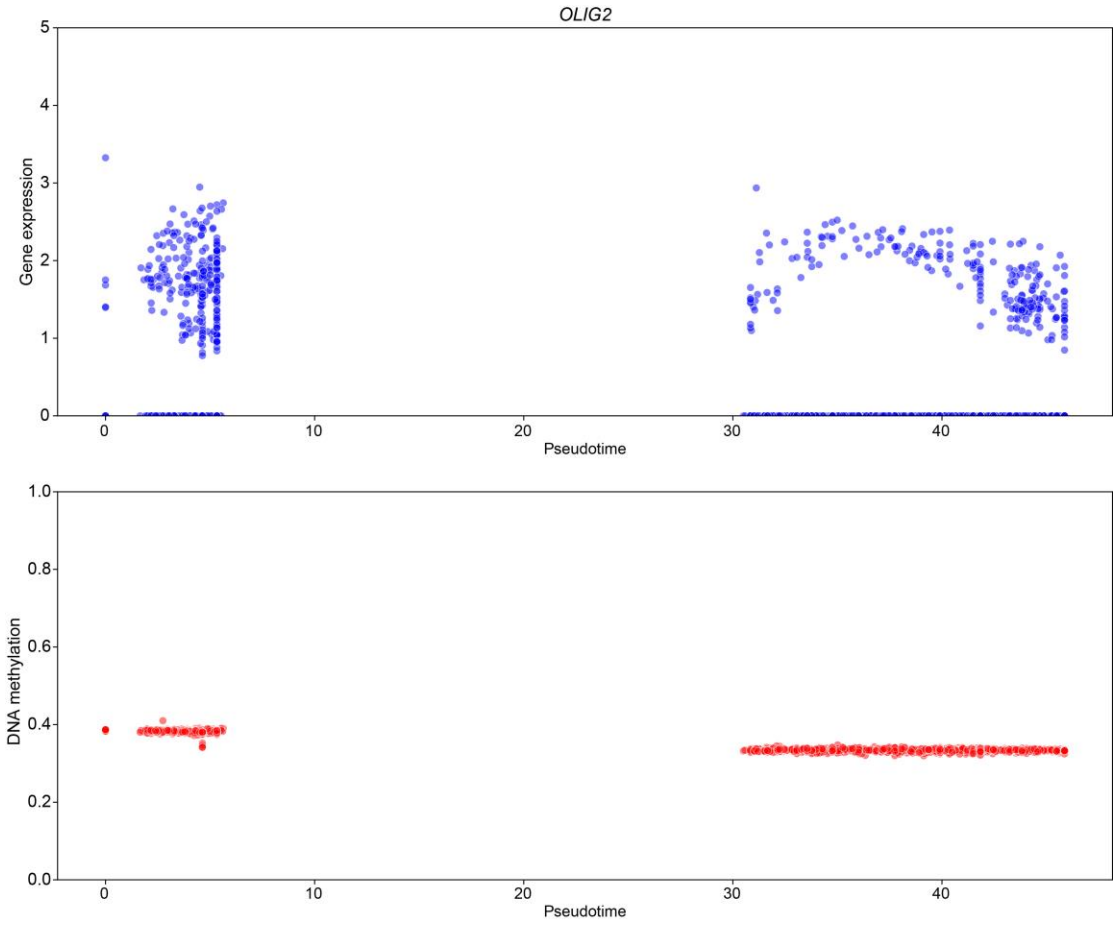

Figure S7. Pseudotime dynamics of *OLIG2* gene expression and DNA methylation along the oligodendrocyte lineage trajectory on the HumanBrainB dataset.

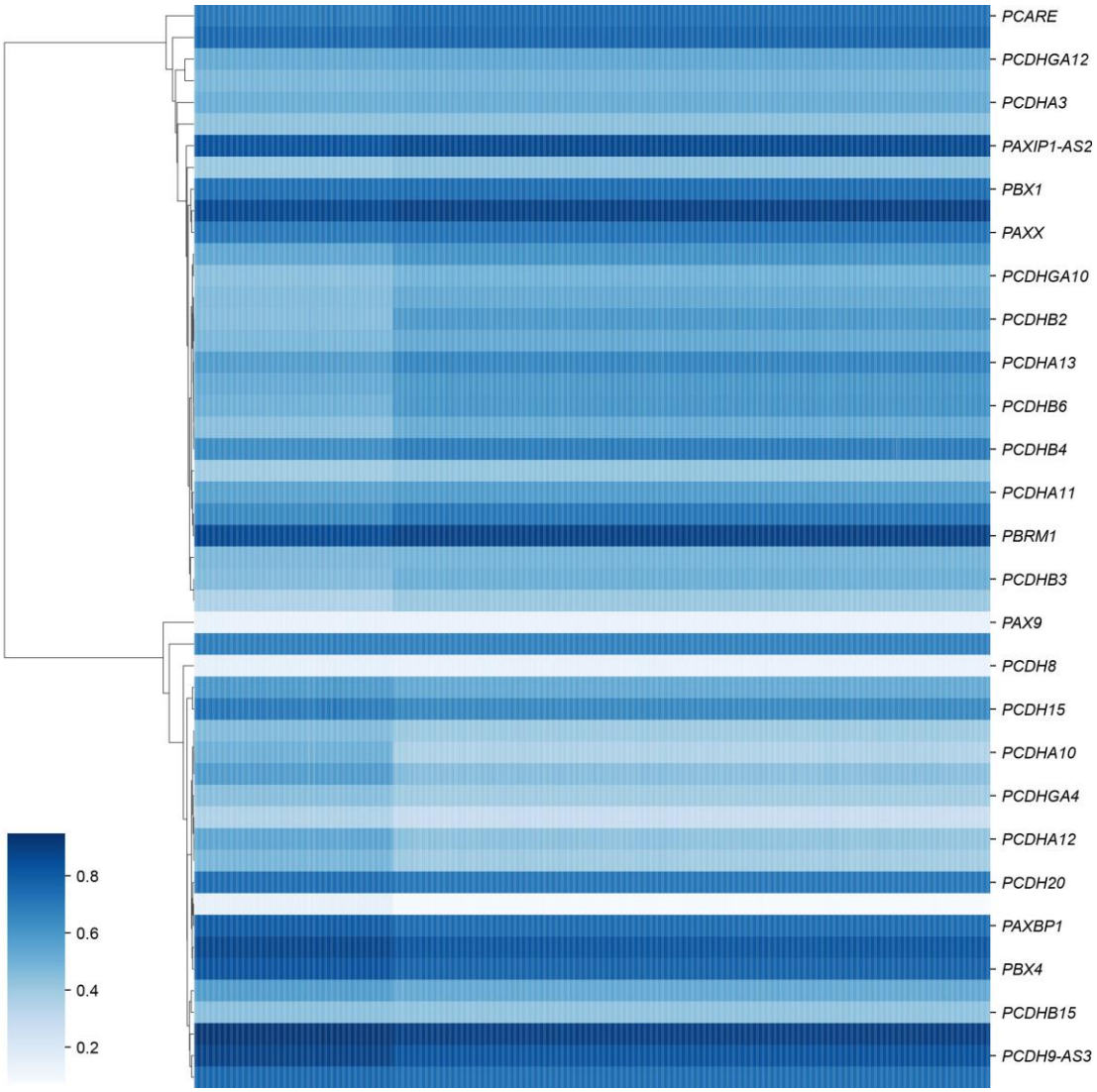

Figure S8. Heatmap of the top 50 genes exhibiting significant promoter DNA methylation changes along the differentiation trajectory on the HumanBrainB dataset.

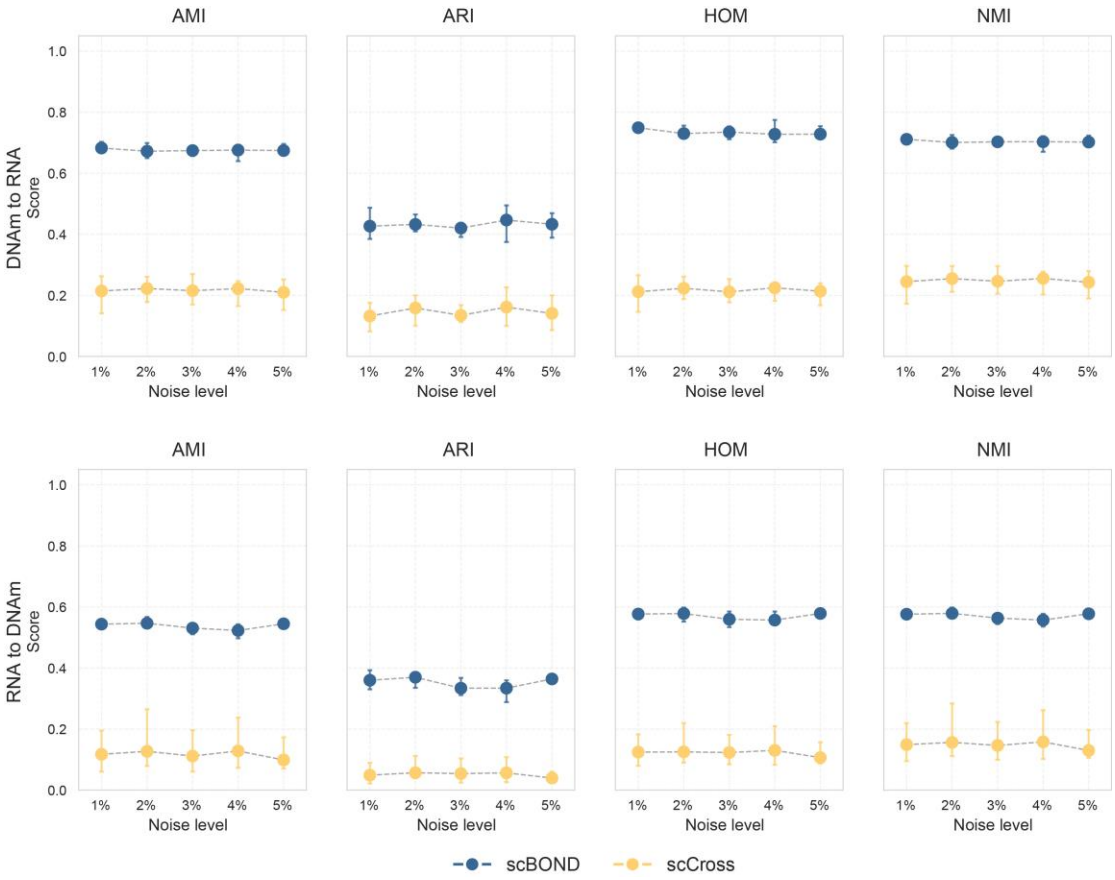

Figure S9. Robustness evaluation of scBOND and baseline methods under increasing RNA noise levels (1%-5%) on the HumanBrainA dataset.

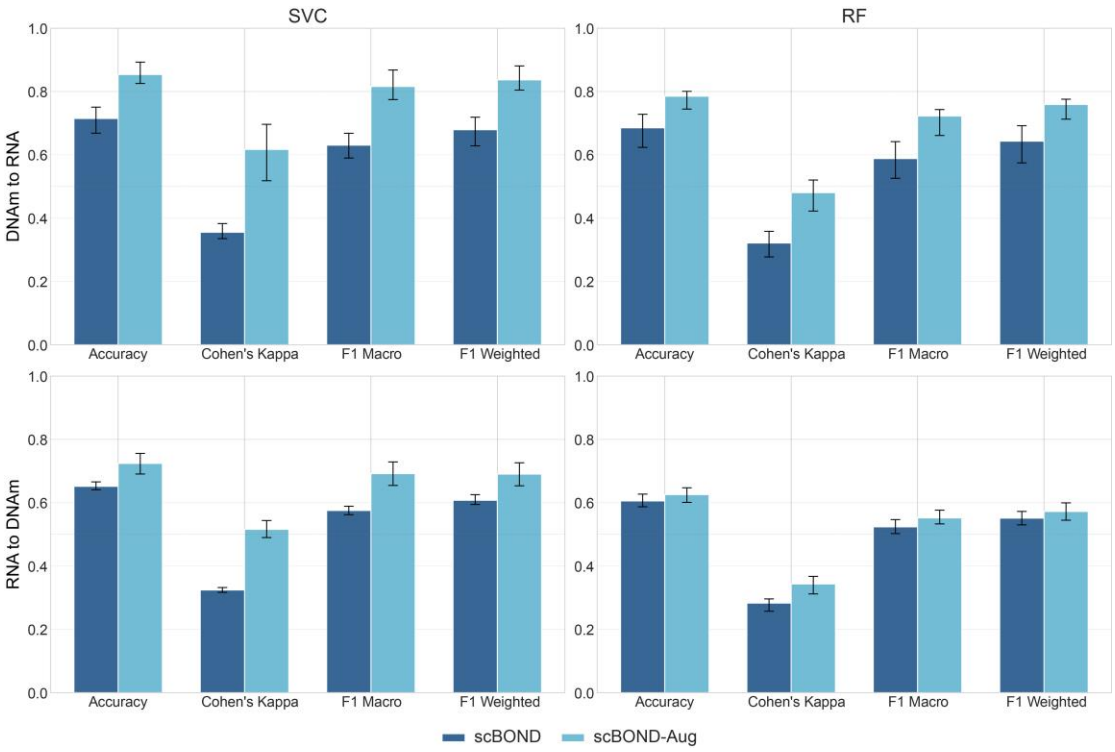

Figure S10. Downstream classification performance of translated profiles generated by scBOND and scBOND-Aug on the HumanBrainA dataset.

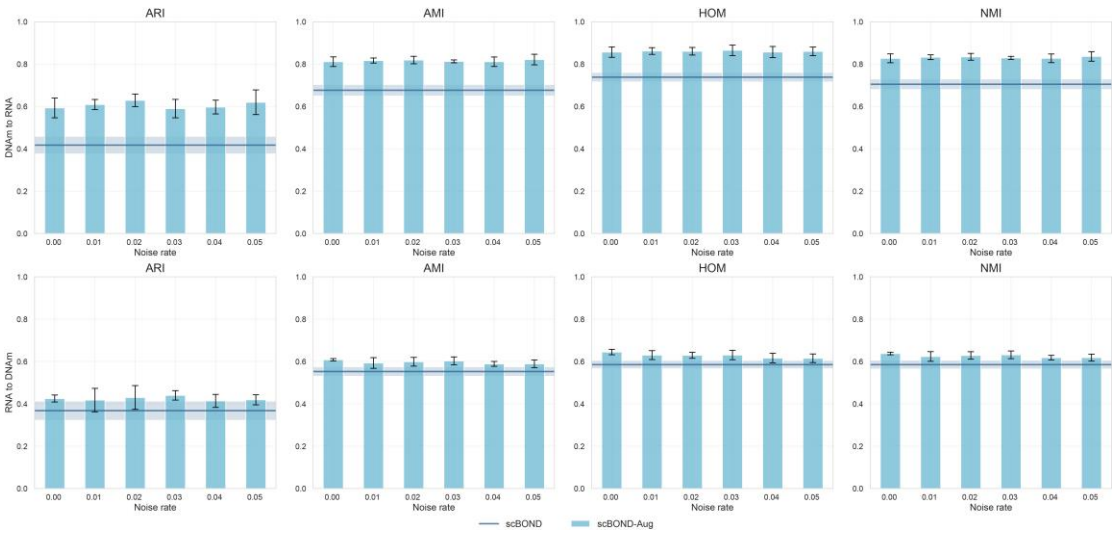

Figure S11. Robustness evaluation of scBOND and scBOND-Aug under increasing cell-type label noise levels (1%-5%) on the HumanBrainA dataset.

References

Abdelaal T, Michielsen L, Cats D, Hoogduin D, Mei H, Reinders MJ, Mahfouz A. 2019. A comparison of automatic cell identification methods for single-cell RNA sequencing data. Genome Biology 20: 194.

Consortium GO. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Research 32: D258-D261.

Danese A, Richter ML, Chaichoompu K, Fischer DS, Theis FJ, Colomé-Tatché M. 2021. EpiScanpy: integrated single-cell epigenomic analysis. Nature Communications 12: 5228.

Finucane HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh P-R, Anttila V, Xu H, Zang C, Farh K. 2015. Partitioning heritability by functional annotation using genome-wide association summary statistics. Nature Genetics 47: 1228-1235.

Huang DW, Sherman BT, Tan Q, Collins JR, Alvord WG, Roayaei J, Stephens R, Baseler MW, Lane HC, Lempicki RA. 2007. The DAVID Gene Functional Classification Tool: a novel biological module-centric algorithm to functionally analyze large gene lists. Genome Bology 8: R183.

Kanehisa M, Goto S. 2000. KEGG: kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27-30.

Slowikowski K, Hu X, Raychaudhuri S. 2014. SNPsea: an algorithm to identify cell types, tissues and pathways affected by risk loci. Bioinformatics 30: 2496-2497.

Traag VA, Waltman L, Van Eck NJ. 2019. From Louvain to Leiden: guaranteeing well-connected communities. Scientific Reports 9: 1-12.

Tu X, Cao Z-J, Xia C-R, Mostafavi S, Gao G. 2022. Cross-linked unified embedding for cross-modality representation learning. In Advances in Neural Information Processing Systems.

Wolf FA, Angerer P, Theis FJ. 2018. SCANPY: large-scale single-cell gene expression data analysis. Genome Biology 19: 15.