nature computational science

Article

https://doi.org/10.1038/s43588-025-00941-3

Denoising spatial epigenomic data via deep matrix factorization

Received: 7 April 2025

Accepted: 24 November 2025

Published online: 13 January 2026

Check for updates

Shuyan Wang[1,2,8], Hao Xu[2,3,8], Junyu Wang[2], Yao Xiao[4,5], Shanghao Dai[2], Junyi Lu[2], Ruoxuan Cao[6], Xuejin Chen[7] & Kun Qu[1,2,3,4]

Spatial epigenomics (SE) technologies profile epigenomic landscapes within intact tissues, preserving spatial context and enabling the study of gene regulatory mechanisms in situ. However, current SE datasets typically suffer from low signal detection, substantial noise and extremely sparse peak matrices, which pose considerable challenges for downstream analysis. Here we introduce SPEED (spatial epigenomic data denoising), a deep matrix factorization framework that leverages atlas-level single-cell epigenomic data and spatial context to impute and denoise SE data. In comprehensive benchmarks on both simulated data and real SE tissue datasets, SPEED outperformed five state-of-the-art methods across diverse tissues and technologies. Moreover, SPEED's denoised outputs facilitated downstream analyses such as differential chromatin accessibility analysis, epigenomic spatial domain identification and gene activity inference. Collectively, our results indicate that SPEED is a generalizable tool for improving data quality and biological insights in SE.

Cellular function within tissues is tightly linked to the spatial organization of cells[1–3]. Recent advances in spatial omics technologies enable the simultaneous profiling of epigenomic, transcriptomic or proteomic landscapes while preserving spatial context, facilitating the construction of spatially resolved single or multi-omics maps[4–11]. Among these, spatial epigenomics (SE) technologies, including spatial ATAC[12], spatial-ATAC-seq[9], epigenomic MERFISH[7], spatial-ATAC-RNA-seq[8], MISAR-seq[13], spatial-CUT&Tag[4], spatial-CUT&Tag-RNA-seq[8], Slide-tags[10] and spatial-Mux-seq[11], enable genome-wide epigenomic profiling at spatial resolution. These technologies have been widely applied to mouse embryonic[4,7–9,11–13] and brain tissues[4,7–9,11], as well as to disease models[10–12,14] and human samples[8,9], providing insights into transcriptional regulatory mechanisms within tissues. However, SE technologies remain challenged by lower signal detection, increased noise and much sparser peak matrices compared with single-cell epigenome sequencing.

For example, the current spatial assays for transposase-accessible chromatin using sequencing (ATAC-seq) methods (including spatial ATAC[12], spatial-ATAC-seq[9] and spatial-ATAC-RNA-seq[8]) generate SE data in which the transcription start site enrichment score and the fraction of reads in peaks are lower than those obtained from single-cell epigenome sequencing of the same tissue[12,15,16] (Supplementary Fig. 1). This limitation persists despite the relatively large spot diameters (20–50 μm), which typically encompass multiple cells. These issues hinder downstream analyses and limit the applicability of SE technologies to complex biological contexts, such as disease tissues.

Several computational approaches have been developed to extract meaningful biological insights from noisy and sparse epigenomic

[1]Department of Oncology, The First Affiliated Hospital of USTC, State Key Laboratory of Eye Health, School of Artificial Intelligence and Data Science, University of Science and Technology of China, Hefei, China. [2]School of Basic Medical Sciences, Division of Life Sciences and Medicine, University of Science and Technology of China, Hefei, China. [3]School of Biomedical Engineering, Suzhou Institute for Advanced Research, University of Science and Technology of China, Suzhou, China. [4]Anhui Province Key Laboratory of Biomedical Imaging and Intelligent Processing, Institute of Artificial Intelligence, Hefei Comprehensive National Science Center, Hefei, China. [5]Institute of Advanced Technology, University of Science and Technology of China, Hefei, China. [6]School of the Gifted Young, University of Science and Technology of China, Hefei, China. [7]MoE Key Laboratory of Brain-Inspired Intelligent Perception and Cognition, University of Science and Technology of China, Hefei, China. [8]These authors contributed equally: Shuyan Wang, Hao Xu. e-mail: qukun@ustc.edu.cn

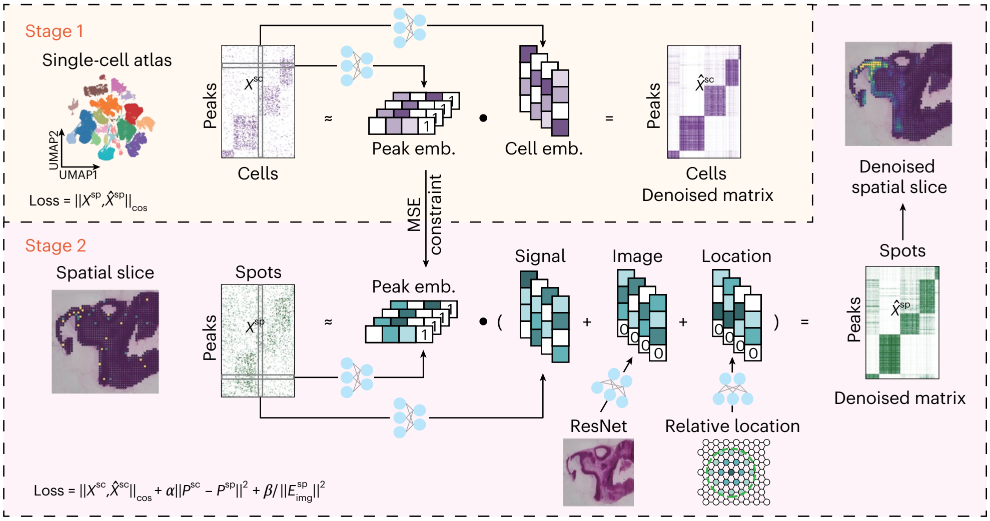

Fig. 1 | Workflow of SPEED. SPEED uses DMF to independently decompose the peak matrices from single-cell and spatial epigenomic data, while imposing a constraint on the similarity between spatial peak embeddings and single-cell peak embeddings to facilitate transfer learning. Spot embeddings are formed by aggregating epigenomic signal embeddings, spatial location embeddings and image embeddings, which are extracted through three distinct neural networks. P^{sc} and P^{sp} respectively denote the peak embeddings for single-cell and spatial data, and E^{sp}_{img} represents the image embedding. The parameter α controls the strength of alignment to single-cell embeddings, and β regulates the contributions of image features in the loss function. MSE, mean squared error; emb., embedding.

data (Supplementary Fig. 2). For instance, scBasset utilizes deep convolutional neural networks to predict chromatin accessibility from DNA sequence features[17]; SCALE integrates a variational autoencoder with a Gaussian mixture model to extract latent representations from single-cell ATAC-seq (scATAC-seq) data[18]; scOpen uses non-negative matrix factorization for scATAC-seq imputation[19]; and cisTopic applies latent Dirichlet allocation to infer topic distributions of cells and genomic regions in scATAC-seq data[20]. However, these methods do not explicitly incorporate spatial information, making them suboptimal for SE data analysis.

spaPeakVAE is a spatially informed method that uses a deep generative spatial variational autoencoder to model spatial ATAC-seq data while capturing spatial correlations between spots[21]. However, it relies on Gaussian process priors to model spatial information, requiring predefined kernel functions to represent correlations across different spatial locations. Furthermore, all these methods use unsupervised learning, which is inherently constrained by the extreme sparsity and noise of SE data.

Concurrently, large-scale single-cell epigenomic atlases have become available, providing comprehensive chromatin accessibility profiles across various tissues, including the mouse embryo[15], mouse brain[22], human brain[23] and human primary tissues[24]. These high-quality single-cell epigenomic datasets offer valuable prior information for modeling SE data, enabling more accurate analyses. To fully leverage this prior information, a computational framework is required that can both denoise SE data and effectively integrate external references. Although originally developed for recommender systems, deep matrix factorization (DMF) captures complex dependencies and recovers meaningful signals from highly sparse and noisy data[25,26], providing a flexible and powerful solution for omics denoising. By incorporating atlas-level single-cell epigenomic data into the DMF framework, peak representations in SE data can be constrained to align with high-quality single-cell priors, thereby improving signal recovery and enhancing downstream analyses.

Here, we propose SPEED (spatial epigenomic data denoising), a DMF framework that leverages atlas-level single-cell data and spatial information for SE data imputation and analysis. Unlike existing approaches, SPEED automatically learns the inherent relationships between peaks from large-scale single-cell epigenomic reference data and transfers this knowledge to the spatial data, effectively mitigating extreme sparsity and noise in SE datasets. In addition, SPEED encodes the spatial arrangement of spots and can incorporate histological image features, enabling it to preserve spatially coherent patterns in the data. Consequently, SPEED achieves better performance in denoising and dimensionality reduction compared with unsupervised methods.

To assess its performance, we systematically evaluated SPEED against five state-of-the-art methods using a range of datasets, including 4 simulated datasets and 14 real tissue sections obtained with spatial ATAC-seq and spatial cleavage under targets and tagmentation (CUT&Tag) technologies. Our results demonstrate that SPEED more accurately imputes spatial chromatin accessibility and histone modification signals while outperforming existing methods in denoising, dimensionality reduction, differential chromatin accessibility analysis and epigenomic spatial domain identification. Collectively, these findings indicate that SPEED is a promising tool for SE data denoising and downstream analysis across multiple SE modalities.

Results

Overview of SPEED

SPEED utilizes atlas-level single-cell epigenomic data as prior knowledge to perform SE data denoising and downstream analysis through a two-step framework. First, SPEED applies DMF to decompose the raw peaks-by-cells matrix from single-cell epigenomic data into two low-dimensional embeddings representing peak and cell features (Fig. 1 and Methods). The reconstructed matrix is obtained by computing the product of these embeddings, with a cosine distance loss function minimizing the distance between the reconstructed and raw matrices to improve signal retention. Next, SPEED applies DMF to the peaks-by-spots matrix from SE data, generating peak and spot embeddings. To transfer learned peak-to-peak relationships from large-scale single-cell data, SPEED constrains the similarity between spatial peak embeddings and single-cell peak embeddings. In parallel, spot embeddings aggregate epigenomic signal embeddings, spatial location embeddings and image embeddings, extracted via three distinct neural networks. The spatial location embeddings are trained using relative spot coordinates, while the optional image embeddings capture additional structural information from high-quality stained images. The loss function consists of three components: the cosine distance loss between the reconstructed and raw matrices, the mean squared error between the peak embeddings of spatial and single-cell data weighted by α, and the L2 regularization constraint on the image embeddings weighted by β. After training, the final product of the spot and peak embeddings represents the denoised SE signal matrix, where spot embeddings integrate epigenomic, spatial and image features.

SPEED distinguishes itself from existing SE data analysis methods in several key aspects: (1) it leverages atlas-level single-cell epigenomic data as a reference to infer peak-to-peak relationships as prior knowledge; (2) it uses cosine distance loss for matrix factorization, enhancing the signal-to-noise ratio of imputed SE data; (3) it adaptively integrates spatial and image features to refine low-dimensional spot embeddings, particularly in structurally similar tissue regions. The learned low-dimensional embeddings capture epigenomic, spatial and image features, enabling robust downstream analyses such as differential chromatin accessibility analysis, epigenomic spatial domain identification and gene activity inference. In addition, SPEED provides a pretrained model for mouse embryo data, leveraging an atlas reference of 312,248 scATAC-seq profiles spanning seven developmental stages[12,15,16,27,28] (Supplementary Data 1). This eliminates the need for de novo training, making SPEED an efficient tool for SE data analysis. Furthermore, its user-friendly design allows researchers to leverage publicly available or self-generated single-cell epigenomic datasets to train models for various tissue types beyond the mouse embryo.

Denoising epigenomic data in simulated datasets

To assess SPEED's denoising performance, we generated simulated SE datasets using scATAC-seq data from 16 tissue types[15]. Cells were aggregated into simulated spots, providing genome-wide chromatin accessibility profiles as the ground truth. We further assigned cells from four randomly picked tissue types to spots in distinct spatial distribution patterns—stripe-like, block-like, circular-like and dispersed—to evaluate SPEED's ability to identify epigenomic spatial domains. To mimic SE data sparsity, we introduced dropouts with an average probability of 90%, generating raw signal matrices for model input (Methods).

Using an independent scATAC-seq dataset as a reference, we applied SPEED to denoise the simulated data (Fig. 2a,b). We applied a commonly used metric, namely the area under the receiver operating characteristic curve (AUROC) to evaluate the accuracy of binary classification models after denoising. A higher AUROC indicates higher similarity between the imputed chromatin accessibility and the ground truth in the simulated data, and thereby better denoising performance (Methods). We found that SPEED outperformed five existing spatial/single-cell ATAC-seq denoising methods (Fig. 2a), achieving higher similarity to the ground truth (Fig. 2c, AUROC per spot/peak = 0.86/0.82) than other methods (AUROC per spot/peak = 0.77–0.86/0.38–0.77).

We then used the adjusted Rand index (ARI) and normalized mutual information (NMI) to assess the similarity between the denoised spatial distribution of spots and the ground truth in simulated data. Higher ARI and NMI values indicate better spatial domain identification by denoising approaches applied to the dropout dataset (Methods). We found that SPEED was the only method capable of recovering the spatial distribution of dispersed spots (Fig. 2b), and its predicted epigenomic spatial domains matched the ground truth more closely than those of other methods (Fig. 2d, ARI/NMI = 0.98/0.97 versus 0.15–0.90/0.20–0.90 for other methods). Finally, we used the Davies–Bouldin index (DBI) and silhouette width (SW) metrics to evaluate the compactness of spots within the same ground truth cluster and the separation of spots between different clusters in the low-dimensional embedding space. Lower DBI and higher SW indicate a clearer distinction of spot clusters in the denoised embeddings and thereby a better identification of spatial domains (Methods). We found that SPEED-derived embeddings provided better separation of ground truth regions compared with other approaches (Fig. 2e, DBI/SW = 0.93/0.85 for SPEED versus DBI/SW = 4.94–1.07/0.51–0.81 for other methods). These results demonstrate SPEED's effectiveness in SE data denoising and epigenomic spatial domain identification.

Recovering tissue-specific chromatin accessibility signals

To evaluate SPEED's performance on real SE data, we applied it to the ATAC modality of the E13 mouse embryo spatial-ATAC-RNA-seq dataset from Zhang et al., a dataset with well-characterized tissue spatial distributions[8]. As a reference, we compiled single-cell or single-nucleus ATAC-seq data from five publicly available datasets, encompassing 312,248 cells across seven developmental stages (embryonic day (E)11.5–E18)[12,15,16,27,28] (Supplementary Data 1). Tissue-specific chromatin accessible sites (TSCAS) were identified through differential accessibility analysis of single-cell or bulk ATAC-seq data from embryonic mouse tissues[15,29]. TSCAS were then defined as the ground truth for evaluating denoising performance in spatial ATAC-seq data, comprising 53,876 TSCAS identified from E13.5 mouse embryo scATAC-seq data and 162,482 TSCAS from bulk ATAC-seq data of the same developmental stage (Supplementary Fig. 3 and Supplementary Data 2). We used fold change (FC) and Moran's *I* to assess the specificity and spatial autocorrelation of all the TSCAS (Methods).

In the raw spatial ATAC-seq data, chromatin accessibility signals at TSCAS appeared diffusely distributed, lacking spatial specificity (Fig. 3a). After applying SPEED, these signals became distinctly localized and continuous in expected tissue regions (Fig. 3a,b and Supplementary Figs. 4 and 5), as reflected by a higher FC of 2.42 versus 1.73 and an increase in Moran's *I* from 0.02 to 0.84. Compared with other methods, SPEED achieved the highest specificity and spatial autocorrelation of ATAC signals at TSCAS (Fig. 3b, FC of 2.42 for SPEED versus 0.51–1.80 for other methods, Moran's *I* = 0.84 versus 0.61–0.81 for other methods), indicating its ability to effectively remove noise while preserving biologically relevant spatial patterns.

To systematically assess SPEED's ability to identify TSCAS, we used E13.5 mouse embryo scATAC-seq and bulk ATAC-seq data (from ENCODE[29]) as the ground truth and computed the Jaccard index (JI) to measure the overlap between differentially accessible chromatin sites identified by different methods and the ground truth datasets (Methods). SPEED exhibited higher similarity to the ground truth (Fig. 3c, JI of bulk/single-cell data, 0.14/0.11) compared with other methods (Supplementary Fig. 6, JI of bulk/single-cell data, 0.06–0.11/0.03–0.06). These results demonstrate that SPEED improves the accuracy of TSCAS identification in spatial ATAC-seq data.

We further investigated whether SPEED preferentially enhances signals at *cis*-regulatory elements compared with other chromatin sites. Using chromatin state annotations from the ENCODE database[30,31] for the E13.5 mouse embryonic forebrain and hindbrain, we categorized all peaks into five major groups: promoter, enhancer, transcription, heterochromatin and others. We then calculated the signal intensity ratios of enhancer and promoter regions relative to other chromatin states before and after denoising (Methods). SPEED-denoised data exhibited the highest signal intensity ratios for promoters and enhancers compared with both raw data and other denoising methods (Fig. 3d, for enhancers, 2.36 for SPEED versus 1.90–2.24 for other methods; for promoters, 4.72 for SPEED versus 3.55–4.43 for other methods). This indicates that SPEED effectively enhances signals at *cis*-regulatory elements.

Identifying epigenomic spatial domains

Identifying epigenomic spatial domains is a critical step in spatial omics data analysis. We used the E13 mouse embryo system as an example, where tissue domain distributions were well characterized by tissue images, including the eye, forebrain, hindbrain, ventricles, limbs and spine, providing a ground truth reference for evaluating spatial domain identification (Fig. 4a). To quantitatively assess the performance of all

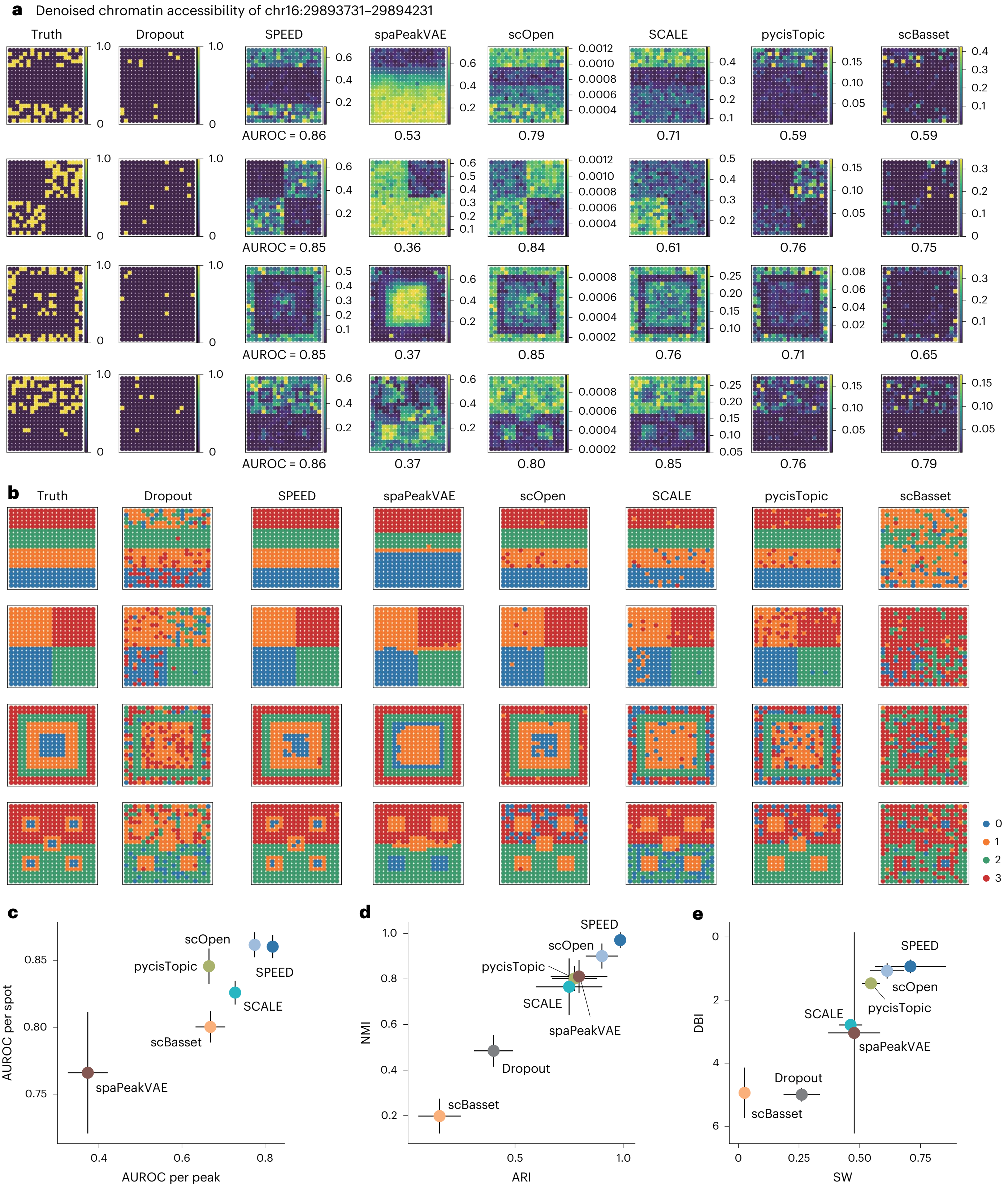

Fig. 2 | Performance on the simulated dataset. a, Spatial distributions of forelimb TSCAS signal obtained from ground truth data, dropout data and denoised data from six methods. **b**, Spatial distributions of epigenomic spatial domains identified by ground truth data, dropout data and denoised data from six methods. **c**, Average AUROC per peak and per spot across six epigenomic denoising methods on simulated datasets. **d**, Average ARI and NMI for dropout data and denoised data across six methods. **e**, Same as **d**, but showing the average DBI and SW. Whiskers, standard errors; $n = 4$ samples in **c**–**e**.

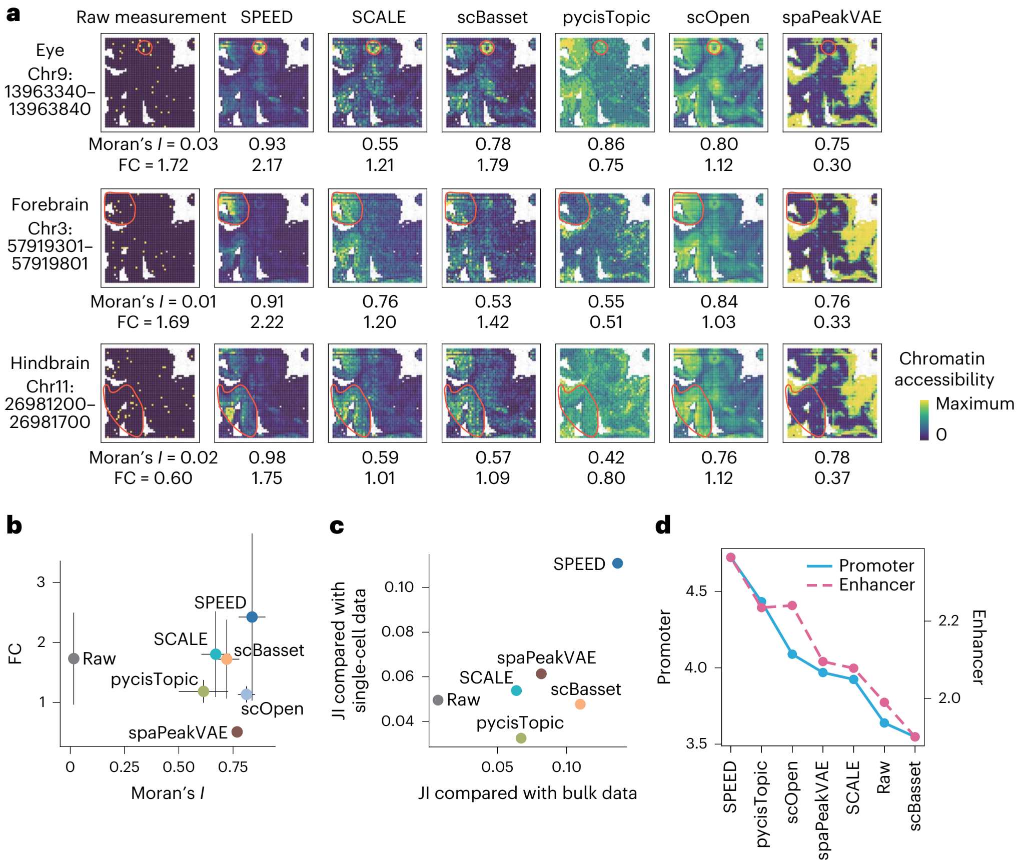

Fig. 3 | Denoising performance on mouse embryo. **a**, Spatial distributions of TSCAS signals for raw measurements and six denoising methods. **b**, Average Moran's *I* and FC comparing TSCAS signals in corresponding tissues with other regions across raw measurements and six methods. Whiskers, 0.5× standard errors; *n* = 41,213/67,250 peaks. **c**, The JI comparing differentially accessible peaks from raw and denoised data with TSCAS identified from scATAC-seq and bulk ATAC-seq. **d**, The ratio of signal intensities in enhancers and promoters relative to other chromatin states for raw measurements and denoised data from six methods. Each dot represents a method or raw measurement; blue and red dots or lines indicate promoter and enhancer signals, respectively. The left and right *y* axes represent promoter and enhancer signal ratios, respectively.

methods in epigenomic spatial domain identification, we also defined two additional ground truth references: (1) spatial domains annotated by joint RNA–ATAC clustering from Zhang et al.[8] (Fig. 4b) and (2) spatial domains inferred from spatial transcriptomics data using SPACEL[32] (Fig. 4c).

Because the raw spatial ATAC-seq data from the E13 mouse embryo spatial-ATAC-RNA-seq dataset are highly noisy and sparse, clustering based on raw data failed to capture the known tissue structures (Fig. 4d). However, when we applied SPEED to the same dataset and clustered the low-dimensional embeddings of spots using the Leiden algorithm, we observed that SPEED successfully identified all tissue regions distinctly (Fig. 4e and Methods), closely aligning with known structures observed in tissue images, as well as the joint RNA–ATAC annotations from Zhang et al. and the spatial domain annotations by SPACEL. These results indicate that the low-dimensional embeddings generated by SPEED, which integrate epigenomic profiles and spatial information, enable precise spatial domain identification.

When we applied all the denoising approaches to the same dataset, we found that only SPEED can successfully recover all tissue structures, particularly in resolving the hindbrain subregion HB2 (JI of 0.64/0.37 for SPEED versus JI of 0.05–0.19/0.04–0.15 for other methods), whereas other methods failed to capture at least one tissue type and often produced mixed or fragmented domains (Supplementary Figs. 7 and 8a). For example, SCALE and scOpen failed to identify the eye and HB2, while spaPeakVAE, pycisTopic and scBasset failed to capture the spine and HB2.

To further quantify SPEED's performance in spatial domain identification compared with other methods, we evaluated four key metrics: ARI, NMI, DBI and SW as above. We found that SPEED-derived spatial domains exhibited stronger agreement with the joint RNA–ATAC annotations (Fig. 4f, ARI/NMI = 0.35/0.49 for SPEED versus ARI/NMI = 0.13–0.31/0.25–0.46 for other methods). In addition, SPEED-derived embeddings yielded the best DBI/SW values compared with other methods, indicating superior tissue-specific separation (Fig. 4g, DBI/SW = 2.59/0.50 for SPEED versus DBI/SW = 7.39–3.03/0.40–0.49 for other methods). These findings were further confirmed using SPACEL-annotated spatial domains as reference, where SPEED consistently achieved the best ARI/NMI and DBI/SW scores (Fig. 4h, ARI/NMI = 0.26/0.50 for SPEED versus ARI/NMI = 0.09–0.18/0.24–0.38 for other methods; Fig. 4i, DBI/SW = 2.45/0.49 for SPEED versus DBI/SW = 6.18–3.33/0.34–0.47 for other methods). These results indicate that, regardless of the ground truth reference used, the epigenomic spatial domains predicted by SPEED consistently align more accurately with known tissue structures.

To further validate SPEED's ability to identify spatial domains across different SE technologies, we applied it to the mouse embryonic brain MISAR-seq dataset from Jiang et al.[13]. This dataset contains spatial ATAC-RNA-seq data across four developmental stages (E11.0, E13.5, E15.5 and E18.5) with eight slices, each paired with hematoxylin and eosin (H&E)-stained images and manual tissue annotations (Fig. 4j and Supplementary Fig. 9). To further leverage H&E images, SPEED extracted image features for each spot and incorporated them into

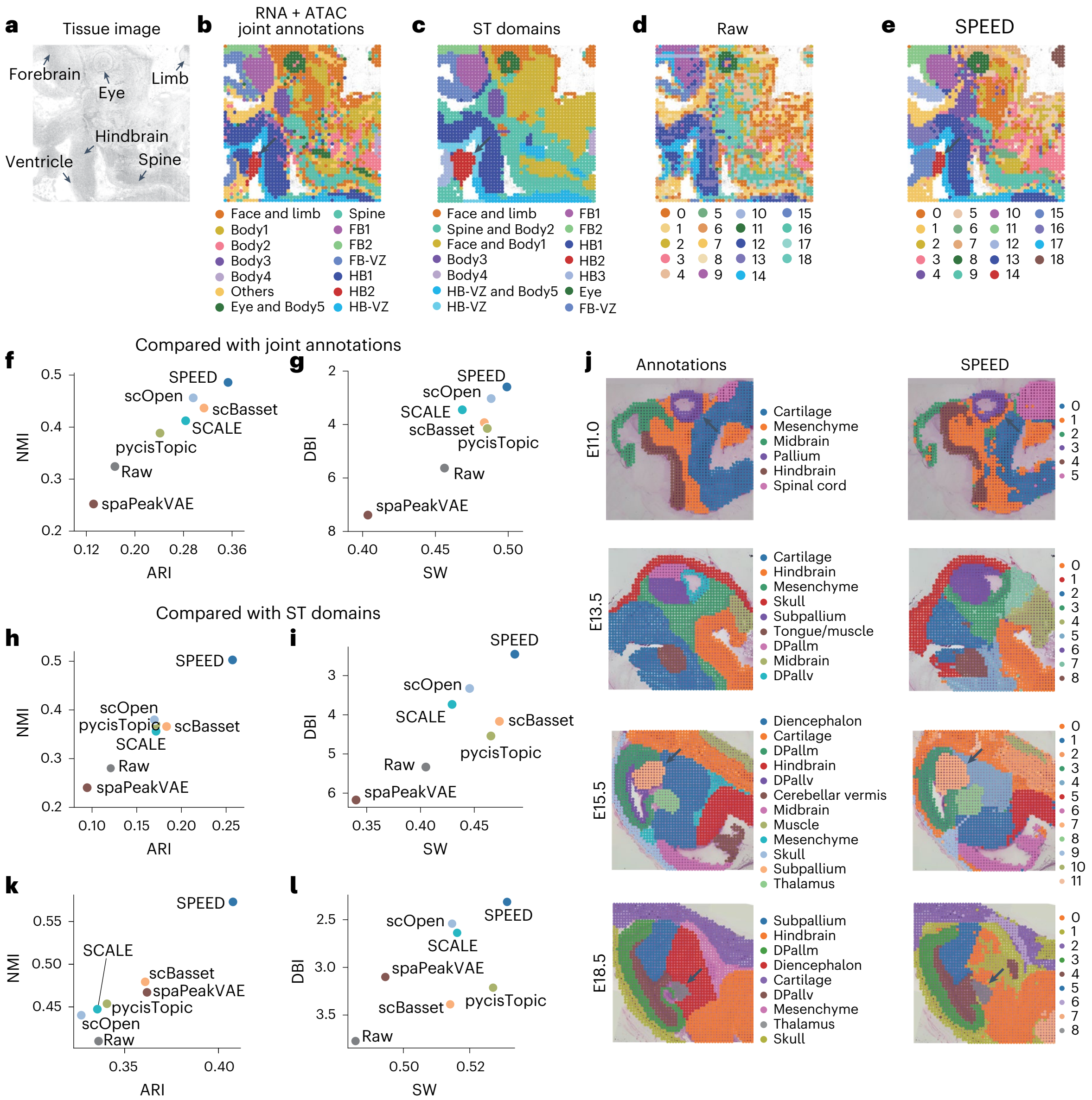

Fig. 4 | Identifying epigenomic spatial domains of mouse embryos. **a**–**e**, Spatial distribution of known tissue and organ locations from the tissue image (**a**), clusters generated from combined ATAC and RNA information provided from original research (**b**), transcriptomic spatial domains identified by SPACEL (**c**), clusters generated using only raw spatial ATAC-seq data (**d**), and epigenomic spatial domains identified by SPEED (**e**). ST, spatial transcriptomic. **f**,**g**, NMI and ARI (**f**), as well as DBI and SW (**g**), comparing epigenomic spatial domains identified from raw and denoised data across six methods, using joint annotations as the ground truth. **h**,**i**, Same as **f** (**h**) and **g** (**i**), but using transcriptomic spatial domains as the ground truth. **j**, Spatial distribution of epigenomic spatial domains annotated with labels from the original study that reference Kaufman's Atlas of Mouse Development and the Allen Brain Atlas (left) and identified by SPEED (right) in mouse embryo brain MISAR-seq data. DPallm, mantle zone of dorsal pallium; DPallv, ventricular zone of dorsal pallium. **k**,**l**, NMI and ARI (**k**), as well as DBI and SW (**l**), comparing the corresponding epigenomic spatial domains identified by raw measurement and denoised data from six methods with manual annotations.

its low-dimensional embeddings, integrating epigenomic, spatial and image features (Methods). Using manual annotations as the ground truth, we found that SPEED effectively distinguished the pallium, subpallium and thalamus regions, whereas other methods failed to accurately identify the boundaries of these regions (Supplementary Fig. 8b, JI of 0.81/0.64/0.77 for SPEED versus JI of 0.18–0.48/0.44–0.50/0.07–0.10 for other methods). Moreover, the SPEED-derived spatial domains exhibited the highest agreement with tissue structures (Fig. 4k, ARI/NMI = 0.41/0.57 for SPEED versus 0.33–0.36/0.44–0.48 for other methods) and more effectively separated distinct regions (Fig. 4l, DBI/SW = 2.31/0.53 for SPEED versus 3.39–2.54/0.49–0.53 for other methods). These results highlight SPEED's robust ability to integrate multimodal spatial information for precise epigenomic spatial domain identification across diverse SE datasets.

To further demonstrate SPEED's utility beyond embryonic tissues, we applied it to three biologically complex datasets: (1) the adult human hippocampus spatial ATAC-RNA-seq dataset from Zhang et al.[8]; (2) the P22 mouse brain dataset from the same study[8]; and (3) the spatial-Mux-seq dataset from the mouse model of neuroinflammation–experimental autoimmune encephalomyelitis (EAE) by Guo et al.[11]. All three datasets reflect greater cellular heterogeneity and complexity compared with embryonic samples.

In the human hippocampus dataset, only SPEED and scOpen successfully distinguished the anatomically annotated choroid plexus

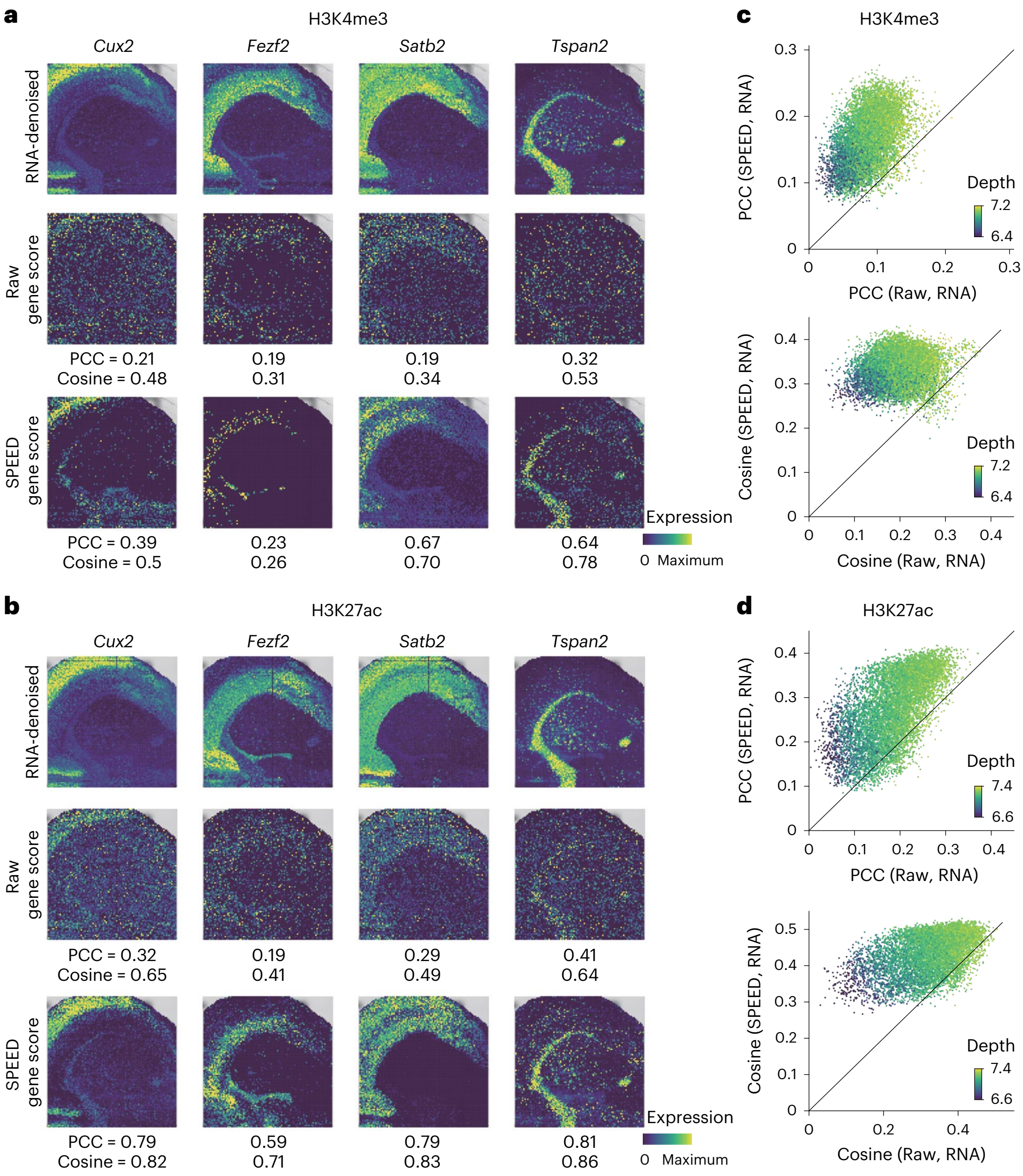

Fig. 5 | Enhancing signals in spatial CUT&Tag data. a,**b**, Spatial distributions of region-specific marker genes. Spots are colored by denoised gene expression using MAGIC, gene scores computed from raw data, and gene scores computed from SPEED-denoised data for H3K4me3 (**a**) and H3K27ac (**b**) in mouse P22 brains. **c**,**d**, PCC and cosine similarity between gene activity scores and denoised gene expression for each spot before (*x* axis) and after (*y* axis) SPEED denoising for H3K4me3 (**c**) and H3K27ac (**d**). Spots are colored according to sequencing depth of spatial transcriptomics data.

and granule cell layer, whereas other methods failed to resolve these structures or produced inaccurate, diffuse boundaries (Supplementary Fig. 10a). In the P22 mouse brain, using the Allen Brain Atlas[33] annotations as the ground truth, SPEED was the only method that resolved four cortical layers, outperforming the original study and other tools that identified two to three layers (Supplementary Fig. 10b,c). In the EAE dataset, SPEED uniquely recovered the lateral ventricle and preserved the cortical layering consistent with the Allen Brain Atlas—structures not discernible in the raw data or with other denoising methods (Supplementary Fig. 10d,e). Together, these results provide strong evidence that SPEED generalizes effectively to adult and disease tissues and supports the discovery of biologically meaningful spatial architecture from complex SE datasets.

Enhancing signals in spatial CUT&Tag data

Spatial CUT&Tag technology enables the study of histone modifications at spatial resolution, yet remains challenged by high noise levels. Because SPEED does not rely on assumptions about spatial distribution, it can be effectively applied to denoise spatial CUT&Tag data.

To evaluate SPEED's performance, we analyzed publicly available P22 mouse brain spatial-CUT&Tag-RNA data[8], which simultaneously profile genome-wide histone modifications and gene expression. We focused on histone H3K4me3 and H3K27ac modifications, which are associated with active chromatin states. Notably, we did not use single-cell CUT&Tag data as a reference due to their higher sparsity compared with spatial data[34,35] (Supplementary Fig. 11 and Methods).

For histone modifications associated with active chromatin, an effective denoising method should enhance the correlation between gene activity inferred from open chromatin regions and gene expression. We computed gene activity scores from both raw and SPEED-denoised data. Across tissue-specific marker genes (*Cux2*, *Fezf2*, *Satb2* and *Tspan2*)[8], SPEED-denoised data exhibited clearer spatial expression patterns corresponding to tissue regions, along with higher correlations with gene expression (Fig. 5a,b).

To quantitatively assess SPEED's impact, we computed the correlation between gene activity scores and gene expression using the Pearson correlation coefficient (PCC) and cosine similarity (Methods). We found that SPEED denoising substantially enhanced the correlations between H3K4me3/H3K27ac modifications and gene expression (Fig. 5c,d and Supplementary Fig. 12). These results demonstrate that SPEED is broadly applicable across different SE data types, providing a versatile framework for signal denoising across diverse SE datasets.

Ablation study of SPEED

To assess the contribution of individual components in SPEED's model design, we conducted ablation studies using the E13 mouse embryo spatial ATAC-seq data from Zhang et al. and the mouse embryonic brain MISAR-seq data from Jiang et al. We compared the full SPEED model against three ablated variant models that lacked either the single-cell reference, spatial information or image features. We found that removing any of the three components resulted in reduced performance (Supplementary Fig. 13a,b). We evaluated multiple image feature extraction networks, including ResNet50[36] and pathology-specific foundation models[37,38], and observed only marginal differences in model performance (Supplementary Fig. 14). By default, SPEED used ResNet50. While excluding image features had little impact on quantitative metrics, it impaired the accurate identification of spatial domains, particularly in regions with well-defined structural boundaries, as the dashed white boxes shown in Supplementary Fig. 13c. Furthermore, incorporating a single-cell reference substantially improved performance over using no reference, whereas supplementing with a large-scale atlas-level reference yielded only modest additional gains beyond the best-performing individual single-cell dataset (Supplementary Fig. 13a,b). These findings confirm that each component of SPEED contributes to its robust performance in optimizing signal denoising and epigenomic spatial domain identification.

Furthermore, we evaluated the impact of different loss functions on model performance. Among models trained with cosine distance loss, binary cross-entropy loss and mean squared error loss, the model with cosine distance loss consistently yielded denoised TSCAS with higher FC and identified epigenomic spatial domains with superior ARI/NMI scores (Supplementary Fig. 15). In addition, to evaluate how reference batch effects and data heterogeneity impact the performance of SPEED, we compared models using single-batch references versus whole-atlas datasets of comparable size. While SPEED exhibited robust performance across all settings, the full atlas consistently yielded the best results (Supplementary Fig. 16). Notably, increasing both the number and diversity of reference cells further enhanced its performance, indicating that broader biological heterogeneity in the reference improves the modeling of peak co-accessibility in SPEED. We also assessed model convergence by tracking training and validation losses, and observed consistent validation loss convergence across all datasets, confirming stable and reliable training (Supplementary Fig. 17).

Discussion

Although SPEED is a promising and generalizable tool for improving data quality and biological insights in SE, it still has certain limitations. Currently, SPEED does not support batch effect correction or cross-slice integration. Incorporating batch correction techniques in SPEED could help to extend its applicability to multislice SE datasets. In terms of model design, SPEED uses a multilayer perceptron network for dimensionality reduction within the DMF framework. In the future, incorporating more advanced dimensionality reduction models—such as graph neural networks or transformer-based methods—into the DMF framework to generate low-dimensional representations, along with mechanisms for automatically learning fusion weights across modalities, may further enhance the performance of SPEED. Besides, SPEED is currently only pretrained on atlas-level single-cell chromatin accessibility datasets of mouse embryo, adult mouse brain and human brain. Expanding the reference to include other tissues, single-cell CUT&Tag data and disease-state single-cell epigenomic datasets, or adopting hybrid strategies that integrate bulk and single-cell references when single-cell data are unavailable, would enhance SPEED's applicability to broader SE data denoising tasks and the analysis of transcriptional regulatory mechanisms in disease contexts.

Although SPEED achieves improved accuracy in signal recovery by integrating spatial information, single-cell references and image features, this integration inevitably comes with increased computational cost. In resource-constrained settings, lighter-weight methods such as scOpen offers advantages in terms of computational efficiency. Nonetheless, this trade-off does not diminish the strengths of SPEED in terms of accuracy and its capacity for novel biological discoveries. Overall, SPEED provides a robust and versatile framework for SE data denoising, facilitating the accurate recovery of chromatin accessibility and histone modification signals across diverse SE datasets.

Methods

SPEED model

For a data matrix $X = \{x_{ij}\}_{N\times M} \in N$, where x_{ij} represents the value of cell or spot i and peak j in the matrix, N represents the number of cells or spots, and M represents the number of peaks, we assume that x_{ij} distributes as a Bernoulli binomial distribution, $x_{ij} \sim \text{Bernoulli}(p_{ij})$. SPEED draws on the DMF framework to model epigenomic data. DMF was initially proposed for recommender systems. Originally developed for recommender systems[25], DMF uses two neural networks to learn low-dimensional representations of the row and column vectors and reconstructs a denser matrix via an inner product or similar combination[26]. This makes DMF naturally suited for denoising sparse epigenomic matrices. In our setting, rows correspond to cells or spots, and columns to peaks. SPEED learns low-dimensional embeddings for each through two fully connected networks. Their inner product yields the predicted probability matrix $\hat{X}$, representing the denoised data, with a Sigmoid activation[39] ensuring non-negativity.

$$\hat{X} = \{\hat{x}_{ij}\}_{N\times M}, x_{ij} \sim \text{Bernoulli}(\hat{x}_{ij}), \tag{1}$$

where $\hat{x}_{ij}$ represents the value of cell or spot i and peak j in the denoised matrix.

The fully connected neural network for extracting low-dimensional embeddings consists of three linear layers. Each output of the linear layer is followed by a LayerNorm layer, Dropout layer and LeakyReLU activation function to produce the final output. For single-cell data, the embeddings $E^{\text{sc}}_{\text{signal}}$ and P^{sc} for each cell and peak are derived by separately reducing the epigenomic profiles via two distinct fully connected neural networks. The reconstruction matrix is given by

$$\hat{X} = \text{Sigmoid}\left(E^{\text{sc}}_{\text{signal}} \cdot P^{\text{sc}} + E^{\text{sc}}_{\text{depth}}\right), \tag{2}$$

where $E^{\text{sc}}_{\text{depth}}$ is learned by the neural network shared with $E^{\text{sc}}_{\text{signal}}$, representing the sequencing depth for each cell.

For spatial data, the embedding $E^{\text{sp}}_{\text{spot}}$ for each spot is obtained by summing the outputs of three distinct fully connected neural networks, which separately process the epigenomic profile embedding $E^{\text{sp}}_{\text{signal}}$, spatial location embedding $E^{\text{sp}}_{\text{loc}}$ and stained image embedding $E^{\text{sp}}_{\text{img}}$. Thus,

$$E^{\text{sp}}_{\text{spot}} = E^{\text{sp}}_{\text{signal}} + E^{\text{sp}}_{\text{loc}} + E^{\text{sp}}_{\text{img}}. \tag{3}$$

Each peak embedding P^{sp} is derived through another fully connected neural network applied to the epigenomic profile. $E^{\text{sp}}_{\text{signal}}$, $E^{\text{sp}}_{\text{loc}}$ and $E^{\text{sp}}_{\text{img}}$ are fused with equal weights into $E^{\text{sp}}_{\text{spot}}$, as the network learns to emphasize or de-emphasize each modality in a data-driven manner.

The reconstruction matrix is

$$\hat{X} = \text{Sigmoid}\left(E^{\text{sp}}_{\text{spot}} \cdot P^{\text{sp}} + E^{\text{sp}}_{\text{depth}}\right), \tag{4}$$

where $E^{\text{sp}}_{\text{depth}}$ is learned by the neural network shared with $E^{\text{sp}}_{\text{signal}}$, representing the sequencing depth for each spot.

The input to the network for extracting cell or spot embeddings is the signal vector $\{\mathbf{x}_{i.}\}$ for each cell or spot *i* across all chromatin accessible peaks and, for extracting peak embeddings, the network inputs the signal vector $\{\mathbf{x}_{.j}\}$ for each peak *j* across all cells. For spatial omics data, H&E image data for each spot were segmented and passed through a ResNet50 (refs. 36,40) network (pretrained on the ImageNet dataset) to extract 2,048-dimensional image features. The image features obtained by the ResNet50 network served as input to the stained image network in SPEED. In addition to the default ResNet50, we also provide options for UNI[37] and Prov-Gigapath[38] models, enabling users to flexibly select different backbones for H&E image feature extraction. The location network receives *k*-order relative locational encodings $\hat{\mathbf{y}}_i$ for each spot as input, where

$$\hat{\mathbf{y}}_i = \left[\hat{d}_{i1}, \hat{d}_{i2}, \ldots, \hat{d}_{iN}\right] \tag{5}$$

$$\hat{d}_{ij} = \begin{cases} 1 + \log\left(\frac{d_{ij}}{d}\right), & \text{if } \frac{d_{ij}}{d} < k \\ 0, & \text{else.} \end{cases} \tag{6}$$

d is the average Euclidean distance between all adjacent spots. $D = \{d_{ij}\}_{N\times N} \in R^{+}$ represents the Euclidean distance matrix among all *N* spots, where d_{ij} represents the Euclidean distance between spot *i* and spot *j*.

Because the number of cells or spots *N* varies greatly across epigenomic datasets, we adapt the size of the multilayer perceptron accordingly to maintain the number of neurons within a practical range of 10^2 to 10^3. In typical spatial epigenomic datasets, *N* ranges from 10^3 to 10^5, while the number of peaks is usually on the order of 10^5, necessitating an adaptive strategy to ensure computational efficiency and stable training. Therefore, for the networks extracting peak and spatial location embeddings, the numbers of neurons in the three layers are $2.5 \times \sqrt{\frac{N+M}{2}}$, $\sqrt{\frac{N+M}{2}}$ and 32, respectively. For the networks extracting cell or spot embeddings, the three layer sizes are $2.5 \times \sqrt{\frac{N+M}{2}}$, $\sqrt{\frac{N+M}{2}}$ and 33, where the first 32 dimensions of the output vector represent $E^{\text{sc}}_{\text{signal}}$ or $E^{\text{sp}}_{\text{signal}}$, and the last dimension encodes the sequencing depth $E^{\text{sc}}_{\text{depth}}$ or $E^{\text{sp}}_{\text{depth}}$. For the stained image embedding network, the three layer sizes are 512, 128 and 32. All modality-specific features—including epigenomic, spatial and image-derived representations—are ultimately embedded into a shared 32-dimensional latent space through the SPEED network, ensuring effective multimodal fusion.

Training processing

To address extreme sparsity in SE data, we use the cosine distance loss function to constrain model training, defined as

$$||X, \hat{X}||_{\text{COS}} = 1 - \frac{1}{2}\text{cosine}(X, \hat{X}) - \frac{1}{2}\text{cosine}\left(X^T, \hat{X}^T\right), \tag{7}$$

where

$$\text{cosine}(X, \hat{X}) = \frac{1}{N}\sum_{i=1}^{N}\text{cosine}(\mathbf{x}_{i.}, \hat{\mathbf{x}}_{i.}) \tag{8}$$

$$\text{cosine}\left(X^T, \hat{X}^T\right) = \frac{1}{M}\sum_{j=1}^{M}\text{cosine}(\mathbf{x}_{.j}, \hat{\mathbf{x}}_{.j}). \tag{9}$$

For single-cell data, the loss function is the cosine distance between the raw matrix and the reconstruction matrix:

$$\text{loss} = ||X, \hat{X}||_{\text{COS}}. \tag{10}$$

The model constrains the similarity between the original input matrix *X* and the reconstructed matrix $\hat{X}$ from the DMF decomposition through the loss function shown in equation (10) to ensure that the DMF decomposition computed by the neural networks is accurate.

For spatial data, SPEED enforces additional constraints to align peak embeddings with single-cell references and regulate the contributions of image features. Therefore, the loss function is

$$\text{loss} = ||X, \hat{X}||_{\text{COS}} + \alpha||P^{\text{sc}} - P^{\text{sp}}||^2 + \beta/||E^{\text{sp}}_{\text{img}}||^2, \tag{11}$$

where *α* controls the strength of alignment to single-cell embeddings, and *β* regulates the contributions of image features. As described above, the model constrains the similarity between the original input matrix *X* and the reconstructed matrix $\hat{X}$ from the DMF decomposition through the loss function shown in equation (11) to ensure that the DMF decomposition computed by the neural networks is accurate.

For spatial datasets lacking high-resolution histological images, we set *β* = 0 during training spatial data. The loss function is as follows:

$$\text{loss} = ||X, \hat{X}||_{\text{COS}} + \alpha||P^{\text{sc}} - P^{\text{sp}}||^2. \tag{12}$$

For spatial datasets lacking matched high-quality single-cell references, we skip the first stage of training on the single-cell dataset and set *α* = 0 during training spatial data. The loss function is as follows:

$$\text{loss} = ||X, \hat{X}||_{\text{COS}} + \beta/||E^{\text{sp}}_{\text{img}}||^2. \tag{13}$$

For spatial datasets lacking both single-cell reference and high-resolution images, we set both *α* = 0 and *β* = 0 during training spatial data. The loss function is as follows:

$$\text{loss} = ||X, \hat{X}||_{\text{COS}}. \tag{14}$$

During training, batches of cells or spots and peaks are sampled separately. Each training epoch iterates through all the cell or spot batches and all the peak batches. The batch size for cells and spots is $2^{\text{int}\left(\log_2 \frac{N}{10}\right)}$, and the batch size for peaks is $2^{\text{int}\left(\log_2 \frac{M}{10}\right)}$. The model parameters are optimized using the Adam optimizer[41] with a learning rate of 0.00001 and weight decay of 0.001. A random subset of 1/6 of the cells or spots and peaks is used as the validation set, with the remaining as the training set. Training continues for up to 500 epochs, with early stopping if the validation loss does not decrease over 30 consecutive epochs.

Binarization method

The reconstruction matrix $\hat{X}$ follows a Bernoulli binomial distribution, so we can calculate the expected proportion of positive signals for each spot and each peak:

$$q_i = \frac{1}{M}\mathrm{E}\left[\sum_{\mathrm{j}} x_{ij}\right] = \frac{1}{M}\sum_{\mathrm{j}}\hat{x}_{ij} \tag{15}$$

$$q_j = \frac{1}{N}\mathrm{E}\left[\sum_{\mathrm{i}} x_{ij}\right] = \frac{1}{N}\sum_{\mathrm{i}}\hat{x}_{ij}. \tag{16}$$

Subsequently, the binarization thresholds for each spot and peak, b_i and b_j, are calculated as

$$b_i = \text{the } q_i\text{th quantile of } \{\mathbf{x}_{i.}\} \tag{17}$$

$$b_j = \text{the } q_j\text{th quantile of } \{\mathbf{x}_{.j}\}. \tag{18}$$

Finally, a binarized matrix X^B is computed from the reconstruction matrix $\hat{X}$ as

$$X^B = \left\{x_{ij}^B\right\}_{N\times N} \tag{19}$$

$$x_{ij}^B = \begin{cases} 1, & \text{if } \hat{x}_{ij} > \frac{b_i+b_j}{2} \\ 0, & \text{else.} \end{cases} \tag{20}$$

Simulation data construction

We sampled scATAC-seq data from Jiang et al.'s E13.5 mouse embryo dataset[15] and synthesized multiple single-cell profiles to generate pseudo-spots. We predefined 400 spots distributed in a 20 × 20 grid. These spots were arranged according to one of four predefined spatial distribution patterns, including stripe-like, block-like, circular-like and dispersed. Each distribution pattern contained four independent spatial regions as ground truth labels, with spots synthesized from scATAC-seq data of different tissue origins. For each spot, the number of cells was simulated using a Gaussian distribution $N(10, 5)$. The spatial region of each spot served as the ground truth for clustering accuracy assessment.

To avoid the introduction of noise from single-cell data, we filtered the signal for each spot based on half of its cell count, setting values below the threshold to zero. The resulting filtered matrix was used as the ground truth for the simulated data to evaluate denoising accuracy. To simulate the sparsity of real spatial ATAC-seq data, random dropouts were introduced into the ground truth matrix by setting certain elements to zero with a defined probability. The dropout probability dp_i for each spot i was generated based on its cell count, calculated as

$$\mathrm{dp}_i = 1 - n_i \times 0.01, \tag{21}$$

where n_i is the cell count of spot i.

When using SPEED to denoise the simulated data, we used only the portion of the mouse embryo sc/snATAC-seq atlas data that does not overlap with the datasets used for simulation as the single-cell reference.

Preprocessing of single-cell and spatial omics data

The raw fastq files of the sc/snATAC-seq data were processed using CellRanger ATAC (v.2.0.0) with the default parameters and aligned to the mm10 reference genome. Peaks were called by MACS2[42] using the 'addReproduciblePeakSet' function in ArchR[43] after merging all single-cell datasets.

For the spatial ATAC-seq data, we used the fragment files provided by the original study, aligned to the mm10 reference genome. We performed denoising and downstream analysis using the peak matrix generated by ArchR, which shared the same peak set with the single-cell reference data.

For the bulk ATAC-seq data, we obtained fragment files and the merged peak set for 16 samples of E13.5 mouse forebrain, midbrain, hindbrain, embryonic facial prominence, limb, liver, heart and neural tube tissues from ENCODE. We then converted the fragments into a samples-by-peaks count matrix.

For the spatial CUT&Tag data, we used the fragment files from the original study, aligned to the mm10 reference genome. We constructed a 500-bp-tiled matrix generated by ArchR for denoising and downstream analysis.

Differential analysis

Differential analysis was performed on the E13.5 mouse embryo scATAC-seq data by grouping cells according to their tissue of origin to identify single cell TSCAS. We normalized the data using the 'RunTFIDF' function in Signac[44], followed by using the 'FindMarkers' function with 'test.use = 'wilcox'' to identify marker peaks for each tissue. Marker peaks were filtered on the basis of $\log_2$FC >1 and adjusted P value <0.01.

Differential analysis was performed on the E13.5 mouse embryo bulk ATAC-seq data based on the tissue origin to identify bulk TSCAS. We used DESeq2[45] to conduct this analysis, selecting marker peaks with $\log_2$FC >1 and adjusted P value <0.01.

For spatial data differential analysis, groups were defined on the basis of joint annotations from the original study, derived from the joint clustering of ATAC and RNA data. To evaluate the similarity of the differential analysis results between the spatial ATAC-seq data and the ground truth (single-cell or bulk data), we applied the both corresponding differential analysis workflows to the spatial data and compared the resulting marker peaks with the respective TSCAS sets. Marker peaks were filtered using $\log_2$FC >0.8 and adjusted P value <0.01.

Identification of epigenomic spatial domains

For each denoising method, we obtained spot embeddings and identified epigenomic spatial domains by applying the Leiden clustering algorithm using the 'scanpy.pp.neighbors' and 'scanpy.tl.leiden' functions in Scanpy[46], with 'random_state=1' to ensure reproducibility. For the raw (undenoised) data, we first performed latent semantic indexing for dimensionality reduction using the 'addIterativeLSI' function in ArchR to generate spot embeddings from fragment files. Subsequently, we applied the 'addClusters' function in ArchR with default parameters to identify epigenomic spatial domains. In addition, for the simulated data before denoising, we used the 'muon.atac.pp.tfidf' and 'muon.atac.tl.lsi' functions in the muon package[47] to derive spot embeddings based on the peak matrix.

Identification of transcriptomic spatial domains

For the E13 mouse embryo spatial-ATAC-RNA-seq data, we used SPACEL[32] to identify transcriptomic spatial domains as the ground truth for the epigenomic spatial domains. The E13 mouse embryo scRNA-seq data from Cao et al.[48] served as the single-cell reference for SPACEL. Specifically, we used the 'SPACEL.Spoint' function for deconvolution and the 'SPACEL.Splane' function to identify transcriptomic spatial domains, with parameters n_neighbors = 4 and k = 1.

Evaluation of denoised spatial ATAC-seq data

AUROC. AUROC is commonly used to evaluate the accuracy of predictions in binary classification tasks. Here, we used it to assess the accuracy of chromatin accessibility predictions by different denoising methods in simulated data. An AUROC score of 1 indicates perfect prediction accuracy, while a score of 0.5 represents random predictions.

Enhancer and promoter signal intensity. We obtained the chromatin states of the E13.5 mouse embryonic forebrain and hindbrain from ENCODE. Then, we annotated the E13 mouse embryo spatial-ATAC-RNA-seq data to the corresponding regions (FB1, FB2 and FB-VZ for forebrain and HB1, HB2 and HB-VZ for hindbrain) and mapped the peaks into five chromatin states (promoter, enhancer, transcription, heterochromatin and others). We scaled the denoised values of each peak to a range of 0–1 based on their 99th and 1st percentiles across different methods, ensuring comparability. Next, we computed the mean signal values of enhancers or promoters in the forebrain and hindbrain. A higher ratio of the mean signal at enhancers or promoters relative to other chromatin states indicates more specific enhancement of signals at *cis*-regulatory element sites.

FC. We calculated the FC of TSCAS signals in specific tissues compared with other tissues in the ATAC modality of E13 mouse embryo spatial-ATAC-RNA-seq data. Tissue annotations were derived from the joint annotations provided by the original study. Specifically, Face and Limb were annotated as Limb; FB1, FB2, FB-VZ, HB1, HB2 and HB-VZ were annotated as Brain; and Eye and Body5 were annotated as Eye. For tissue t, the FC for each marker m_t is calculated as

$$\text{Fold change}\left(m_t\right) = \frac{\max\limits_{T}\left(\sum_{i\in\{S_l^T\}}\hat{x}_{\mathrm{im}_t}/n_T\right)}{\max\limits_{T^{\mathrm{O}}}\left(\sum_{i\in\{S_l^{\mathrm{O}}\}}\hat{x}_{\mathrm{im}_t}/n_{T^{\mathrm{O}}}\right)}, \tag{22}$$

where $\hat{X} = \{\hat{x}_{ij}\}_{N\times M}$ represents the denoised matrix. For tissue t, T represents each of the joint annotations corresponding to t, with the set of included spots denoted as $S_l^T, l = 1, 2, \ldots, n_T$; and T^{O} represents each of the other joint annotations, with the set of included spots denoted as $S_l^{\mathrm{O}}, l = 1, 2, \ldots, n_{T^{\mathrm{O}}}$. Higher FC values indicate greater tissue specificity of TSCAS.

Moran's *I*. We calculated Moran's *I* to assess the spatial autocorrelation of denoised TSCAS signals in the ATAC modality of E13 mouse embryo spatial-ATAC-RNA-seq data. For each spot, we identified its four nearest neighbors within the same joint annotation from the original study. Then, we computed Moran's *I* for each peak using 'scanpy.metrics.morans_i'. Higher Moran's *I* indicates higher spatial autocorrelation.

Evaluation of spatial ATAC-seq data differential analysis results

JI. We used the JI to calculate the similarity between the differential accessible peaks obtained from different denoising methods and the TSCAS. For each joint annotation a from the original study, the set of differential accessible peaks are denoted as $\tilde{P}_a$, and for each tissue t in single-cell or bulk data, the TSCAS is denoted as P_t. The JI between them is defined as

$$\mathrm{JI}\left(a,t\right) = \frac{|\tilde{P}_a \cap P_t|}{|\tilde{P}_a \cup P_t|}. \tag{23}$$

When comparing different methods, we calculate the average of the maximum JI values for each tissue as the overall JI for each method:

$$\mathrm{JI} = \frac{1}{T}\sum_t \max{}_a\left(\mathrm{JI}\left(a,t\right)\right), \tag{24}$$

where T is the number of tissues. A higher JI indicates a greater similarity between the differentially accessible peaks and TSCAS. Notably, because scOpen denoises normalized peak matrices rather than raw peak matrices, it is not suitable for differential analysis using the same workflow as applied to raw single-cell and bulk data, and therefore cannot be evaluated using JI.

Evaluation of spatial ATAC-seq dimensionality reduction results

We evaluated the separation of ground truth labels in the low-dimensional latent space generated by different methods using the DBI and SW.

DBI. Let K denote the number of ground truth labels. s_i represents the average distance from all points in the ith cluster to its center in the low-dimensional latent space, and d_{ij} represents the distance between the centers of the ith and jth clusters. The DBI is then defined as

$$\mathrm{DBI} = \frac{1}{K}\sum_{i=1}^{K}\max_{j\neq i} R_{ij}, \tag{25}$$

where:

$$R_{ij} = \frac{s_i + s_j}{d_{ij}}. \tag{26}$$

A lower DBI indicates a greater separation of the ground truth labels in the low-dimensional latent space.

SW. Let K denote the number of ground truth labels. $d\left(k,l\right)$ represents the distance between spot k and l in the latent space. For each spot $k \in C_K$, let

$$a\left(k\right) = \frac{1}{|C_K| - 1}\sum_{l\in C_K, l\neq k} d\left(k,l\right) \tag{27}$$

$$b\left(k\right) = \min_{L\neq K}\frac{1}{|C_L|}\sum_{l\in C_L} d\left(k,l\right), \tag{28}$$

where $|C_K|$ is the number of spots belonging to cluster C_K. The SW is computed as

$$\text{Silhouette} = \sum_{k=1}^{K}\frac{b\left(k\right) - a\left(k\right)}{\max\left[a\left(k\right), b\left(k\right)\right]}. \tag{29}$$

The score is then scaled between 0 and 1, following the approach used in scib[49]. A higher SW indicates a greater separation of the ground truth labels in the low-dimensional latent space.

Evaluation of epigenomic spatial domain identification

Ground truth annotations used for evaluation. For the E13 mouse embryo spatial ATAC-RNA-seq data, we defined two ground truth references: (1) clusters generated by joint clustering of spatial ATAC-RNA data from Zhang et al. and (2) spatial domains inferred from spatial transcriptomics data using SPACEL. According to the anatomical annotations from the original study, we mapped these clusters and spatial domains to forebrain, hindbrain, eye, limb, facial and body, and so on.

For the mouse embryo MISAR-seq data, we used the manual annotations of each tissue (from the original study, referencing Kaufman's Atlas of Mouse Development[50] and the Allen Brain Atlas[33]) as the ground truth.

We used the NMI and ARI to evaluate the concordance between the epigenomic spatial domains identified by different methods and the ground truth labels. The number of domains generated by each method was consistent with the number of ground truth labels.

NMI. For ground truth labels T and epigenomic spatial domains A:

$$\mathrm{NMI}\left(T,A\right) = \frac{2\times I\left(T,A\right)}{H\left(T\right) + H\left(A\right)}, \tag{30}$$

where $I(.,.)$ represents mutual information and $H(.)$ represents entropy. A larger NMI indicates better alignment between the clustering and the ground truth labels.

ARI. The ARI is calculated as

$$\mathrm{ARI} = \frac{\mathrm{RI} - E\left(\mathrm{RI}\right)}{\max\left(\mathrm{RI}\right) - E\left(\mathrm{RI}\right)}, \tag{31}$$

where

$$\mathrm{RI} = \frac{\mathrm{tp} + \mathrm{tn}}{\binom{N}{2}}. \tag{32}$$

tp represents the number of true positives, and tn represents the number of true negatives. A higher ARI indicates a better match between the epigenomic spatial domains and the ground truth labels.

Evaluation of spatial CUT&Tag-RNA data denoising results

We assessed the accuracy of denoising results by calculating the similarity between gene activity scores obtained from denoised spatial CUT&Tag data and gene expression. Gene expression data are denoised using MAGIC[51]. Specifically, we processed the raw fragment files and the binarized denoised outputs from SPEED using ArchR to obtain gene activity scores for the raw and data denoised by SPEED. Subsequently, we assessed the correlation between gene activity scores and gene expression at each spot using the PCC and cosine similarity.

$$\mathrm{PCC}\,(X,Y)=\frac{\mathrm{cov}\,(X,Y)}{\sigma_X\sigma_Y} \tag{33}$$

$$\mathrm{Cosine}\,(X,Y)=\frac{X\cdot Y}{||X||\cdot||Y||}. \tag{34}$$

Hyperparameter settings for benchmarking methods

The selection of hyperparameters for each method followed the official tutorials and codes provided by the respective authors: spaPeakVAE[52], scBasset[53], pycisTopic[54], scOpen[55] and SCALE[56].

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All SE, single-cell RNA-seq and scATAC-seq datasets used in this study can be downloaded from public websites or databases: E11.5, E12.5 and E13.5 mouse embryo scATAC-seq data from 16 tissues at https://ngdc.cncb.ac.cn/gsa/browse/CRA003910 (ref. 15). E12.5, E13.5 and E14.5 embryonic mouse cerebellum snATAC-seq data are available in the GEO database under accession GSE178546 (ref. 28). E17.5 embryonic mouse heart scATAC-seq data are available in the GEO database under accession GSE190977 (ref. 27). E12.5, E13.5 and E15.5 mouse embryo snATAC-seq data are available in the GEO database under accession GSE214991 (ref. 12). Three samples of E18 mouse embryo brain snATAC-seq data are available at https://www.10xgenomics.com/datasets (ref. 16). Mouse embryo scRNA-seq data are available in the GEO database under accession GSE119945 (ref. 48). Human brain scATAC-seq data are available in the GEO database under accession GSE147672 (ref. 57). Adult mouse brain scATAC-seq data are available in the GEO database under accession GSE246791 (ref. 22). E13 mouse embryo spatial-ATAC-RNA-seq data are available in the GEO database under accession GSE205055 (ref. 8). E11–E18.5 mouse embryo MISAR-seq data are available at https://www.biosino.org/node/project/detail/OEP003285 (ref. 13). P22 mouse brain Spatial-CUT&Tag-RNA-seq data are available in the GEO database under accession GSE205055 (ref. 8). Human hippocampus spatial-ATAC-RNA-seq data are available in the GEO database under accession GSE205055 (ref. 8). P22 mouse brain spatial-ATAC-RNA-seq data are available in the GEO database under accession GSE205055 (ref. 8). EAE mouse brain spatial-Mux-seq data are available in the GEO database under accession GSE263333 (ref. 11). E13.5 mouse embryonic forebrain, hindbrain, midbrain and limb bulk ATAC-seq data from ENCODE are available at https://www.encodeproject.org (ref. 29). Chromatin state annotations for the E13.5 mouse embryonic forebrain and hindbrain are available at https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=2471038369_O34GqlYujAEy04rHeqMnjX560AHY&g=encode3RenChromHmm (refs. 30,31). Source data are provided with this paper.

Code availability

The open-source package of SPEED is available via GitHub at https://github.com/QuKunLab/SPEED. All codes and scripts used for the analyses and figure plotting in this study are available via Zenodo at https://doi.org/10.5281/zenodo.14948507 (ref. 58).

References

1. Bergmann, S. et al. Spatial profiling of early primate gastrulation in utero. *Nature* **609**, 136–143 (2022).
2. Chen, A. et al. Single-cell spatial transcriptome reveals cell-type organization in the macaque cortex. *Cell* **186**, 3726–3743 (2023).
3. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).
4. Deng, Y. et al. Spatial-CUT&Tag: spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681–686 (2022).
5. Unterauer, E. M. et al. Spatial proteomics in neurons at single-protein resolution. *Cell* **187**, 1785–1800 (2024).
6. Liu, Y. et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. *Nat. Biotechnol.* **41**, 1405–1409 (2023).
7. Lu, T., Ang, C. E. & Zhuang, X. Spatially resolved epigenomic profiling of single cells in complex tissues. *Cell* **185**, 4448–4464 (2022).
8. Zhang, D. et al. Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113–122 (2023).
9. Deng, Y. et al. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375–383 (2022).
10. Russell, A. J. C. et al. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature* **625**, 101–109 (2024).
11. Guo, P. F. et al. Multiplexed spatial mapping of chromatin features, transcriptome and proteins in tissues. *Nat. Methods* **22**, 520–529 (2025).
12. Llorens-Bobadilla, E. et al. Solid-phase capture and profiling of open chromatin by spatial ATAC. *Nat. Biotechnol.* **41**, 1085–1088 (2023).
13. Jiang, F. et al. Simultaneous profiling of spatial gene expression and chromatin accessibility during mouse brain development. *Nat. Methods* **20**, 1048–1057 (2023).
14. Kong, D. et al. Spatial profiling of chromatin accessibility reveals alteration of glial cells in Alzheimer's disease mouse brain. Preprint at *bioRxiv* https://doi.org/10.1101/2025.05.01.651759 (2025).
15. Jiang, S. et al. Single-cell chromatin accessibility and transcriptome atlas of mouse embryos. *Cell Rep.* **42**, 112210 (2023).
16. 10x Genomics Datasets. *10X Genomics* https://www.10xgenomics.com/resources/datasets (2019).
17. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).
18. Xiong, L. et al. SCALE method for single-cell ATAC-seq analysis via latent feature extraction. *Nat. Commun.* **10**, 4576 (2019).
19. Li, Z. et al. Chromatin-accessibility estimation from single-cell ATAC-seq data with scOpen. *Nat. Commun.* **12**, 6386 (2021).
20. Bravo Gonzalez-Blas, C. et al. cisTopic: *cis*-regulatory topic modeling on single-cell ATAC-seq data. *Nat. Methods* **16**, 397–400 (2019).
21. Tian, T., Zhang, J., Lin, X., Wei, Z. & Hakonarson, H. Dependency-aware deep generative models for multitasking analysis of spatial omics data. *Nat. Methods* **21**, 1501–1513 (2024).
22. Zu, S. et al. Single-cell analysis of chromatin accessibility in the adult mouse brain. *Nature* **624**, 378–389 (2023).
23. Li, Y. E. et al. A comparative atlas of single-cell chromatin accessibility in the human brain. *Science* **382**, eadf7044 (2023).
24. Zhang, K. et al. A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985–6001 (2021).
25. Xue, H. J., Dai, X. Y., Zhang, J. B., Huang, S. J. & Chen, J. J. Deep matrix factorization models for recommender systems. In *Proc. 26th International Joint Conference on Artificial Intelligence* (ed. Sierra, C.) 3203–3209 (IJCAI, 2017).
26. Yi, B. L. et al. Deep matrix factorization with implicit feedback embedding for recommendation system. *IEEE Trans. Ind. Inf.* **15**, 4591–4601 (2019).

27. Yamada, S. et al. TEAD1 trapping by the Q353R-Lamin A/C causes dilated cardiomyopathy. *Sci. Adv.* **9**, eade7047 (2023).
28. Khouri-Farah, N., Guo, Q., Morgan, K., Shin, J. & Li, J. Y. H. Integrated single-cell transcriptomic and epigenetic study of cell state transition and lineage commitment in embryonic mouse cerebellum. *Sci. Adv.* **8**, eabl9156 (2022).
29. Encode Project Consortium An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
30. Gorkin, D. U. et al. An atlas of dynamic chromatin landscapes in mouse fetal development. *Nature* **583**, 744–751 (2020).
31. Perez, G. et al. The UCSC Genome Browser database: 2025 update. *Nucleic Acids Res.* **53**, D1243–D1249 (2025).
32. Xu, H. et al. SPACEL: deep learning-based characterization of spatial transcriptome architectures. *Nat. Commun.* **14**, 7603 (2023).
33. Harris, J. A. et al. Hierarchical organization of cortical and thalamic connectivity. *Nature* **575**, 195–202 (2019).
34. Bartosovic, M. & Castelo-Branco, G. Multimodal chromatin profiling using nanobody-based single-cell CUT&Tag. *Nat. Biotechnol.* **41**, 794–805 (2023).
35. Bartosovic, M., Kabbe, M. & Castelo-Branco, G. Single-cell CUT&Tag profiles histone modifications and transcription factors in complex tissues. *Nat. Biotechnol.* **39**, 825–835 (2021).
36. He, K. M., Zhang, X. Y., Ren, S. Q. & Sun, J. Deep residual learning for image recognition. In *Proc. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (IEEE, 2016).
37. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
38. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
39. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
40. Wightman, R. PyTorch image models. *GitHub* https://github.com/rwightman/pytorch-image-models (2019).
41. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations* https://arxiv.org/abs/1412.6980 (2015).
42. Zhang, Y. et al. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
43. Granja, J. M. et al. ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
44. Stuart, T., Srivastava, A., Madad, S., Lareau, C. A. & Satija, R. Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
45. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
46. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
47. Bredikhin, D., Kats, I. & Stegle, O. MUON: multimodal omics analysis framework. *Genome Biol.* **23**, 42 (2022).
48. Cao, J. et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
49. Luecken, M. D. et al. Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
50. Kaufman, M. H. *The Atlas of Mouse Development* (Academic Press, 1992).
51. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716–729 (2018).
52. spVAE. *GitHub* https://github.com/ttgump/spaVAE/blob/main/src/spaPeakVAE/run_spaPeakVAE.py (2024).
53. scBasset. *GitHub* https://github.com/calico/scBasset/blob/main/README.md (2022).
54. Using pycisTopic on human cerebellum single-cell multiome data. *pycisTopic* https://pycistopic.readthedocs.io/en/latest/notebooks/human_cerebellum.html (2022).
55. scopen. *GitHub* https://github.com/CostaLab/scopen/blob/master/README.md (2021).
56. SCALE. *GitHub* https://github.com/jsxlei/SCALE/blob/master/README.md (2019).
57. Corces, M. R. et al. Single-cell epigenomic analyses implicate candidate causal variants at inherited risk loci for Alzheimer's and Parkinson's diseases. *Nat. Genet.* **52**, 1158–1168 (2020).
58. Wang, S. Scripts and data for paper titled "Denoising spatial epigenomic data via deep matrix factorization". *Zenodo* https://doi.org/10.5281/zenodo.14948507 (2025).

Acknowledgements

This work was supported by the National Natural Science Foundation of China grants (grant nos T2125012 and 92574202 to K.Q.; 323B2014 to H.X.), the National Key R&D Program of China (grant no. 2022YFA1303200 to K.Q.), Strategic Priority Research Program of Chinese Academy of Sciences (grant no. XDB0940301 to K.Q.) and USTC Research Funds of the Double First-Class Initiative (grant nos YD9100002026 and YD9100002032 to K.Q.). We thank the USTC supercomputing center and the School of Life Science Bioinformatics Center for providing computing resources for this project.

Author contributions

K.Q. conceived the project. S.W. and H.X. designed the framework and performed data analysis with help from J.W., Y.X., S.D., J.L., R.C. and X.C. K.Q., S.W. and H.X. wrote the paper with input from all authors. K.Q. supervised the entire project. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s43588-025-00941-3.

Correspondence and requests for materials should be addressed to Kun Qu.

Peer review information *Nature Computational Science* thanks Chaoyong Yang and Zexian Zeng for their contribution to the peer review of this work. Peer reviewer reports are available. Primary Handling Editors: Michelle Badri and Ananya Rastogi, in collaboration with the *Nature Computational Science* team.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

nature portfolio

Corresponding author(s): Kun Qu

Last updated by author(s): Nov 28, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
□	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
□	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
□	☒	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
□	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
□	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
□	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for data collection.
Data analysis	We used Cell Ranger ATAC (v2.0.0) and ArchR (v1.0.2) for data preprocessing, and used Scanpy (v1.9.3), Muon (v0.1.3), and SPACEL (v1.1.7) to assist with result evaluation. We compared the performance of the SPEED with 5 denoising methods: scOpen (v0.1.7), SCALE (v1.2.1), scBasset (v0.1), spaPeakVAE (Github commit 3673cad), pycisTopic (v1.0.3) The SPEED open source package is available at a GitHub repository: https://github.com/QuKunLab/SPEED. We uploaded the code and scripts used for the analysis and figure plotting to a public Zenodo repository (https://doi.org/10.5281/zenodo.14948507).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All spatial epigenomics, single-cell RNA-seq, and single-cell ATAC-seq datasets used in this study can be downloaded from public websites and/or databases:
(1) E11.5, E12.5, and E13.5 mouse embryo scATAC-seq data from 16 tissues at https://ngdc.cncb.ac.cn/gsa/browse/CRA003910.
(2) E12.5, E13.5, and E14.5 embryonic mouse cerebellum snATAC-seq data are available in the GEO database under accession GSE178546.
(3) E17.5 embryonic mouse heart scATAC-seq data is available in the GEO database under accession GSE190977.
(4) E12.5, E13.5, and E15.5 mouse embryo snATAC-seq data are available in the GEO database under accession GSE214991.
(5) Three samples of E18 mouse embryo brain snATAC-seq data are available at https://www.10xgenomics.com/datasets.
(6) Adult mouse brain scATAC-seq data is available in the GEO database under accession GSE246791.
(7) Human brain scATAC-seq data is available in the GEO database under accession GSE147672.
(8) Mouse embryo scRNA-seq data is available in the GEO database under accession GSE119945.
(9) E13 mouse embryo spatial-ATAC-RNA-seq data is available in the GEO database under accession GSE205055.
(10) P22 mouse brain spatial-ATAC-RNA-seq data is available in the GEO database under accession GSE205055.
(11) Human hippocampus spatial-ATAC-RNA-seq data is available in the GEO database under accession GSE205055.
(12) EAE mouse brain spatial-Mux-seq data is available in the GEO database under accession GSE263333.
(13) E11-E18.5 mouse embryo MISAR-seq data are available at https://www.biosino.org/node/project/detail/OEP003285.
(14) P22 mouse brain Spatial-CUT&Tag-RNA-seq data is available in the GEO database under accession GSE205055.
(15) E13.5 mouse embryonic forebrain, hindbrain, midbrain, and limb bulk ATAC-seq data from ENCODE are available at https://www.encodeproject.org.
(16) Chromatin state annotations for the E13.5 mouse embryonic forebrain and hindbrain are available at https://genome.ucsc.edu/cgi-bin/hgTrackUi?hgsid=2471038369_O34GqlYujAEy04rHeqMnjX560AHY&g=encode3RenChromHmm.
We also provide a public Zenodo repository for users to download all the above datasets (https://doi.org/10.5281/zenodo.14948507).

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences □ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size

We used 6 spatial epigenomics datasets (14 slices), 7 scATAC-seq datasets, and 1 scRNA-seq dataset from published studies. The details of these datasets are listed as follows:
(1) 1 slice from E13 mouse embryo spatial-ATAC-RNA-seq dataset: 2,187 spots;
(2) 8 slices from E13 mouse embryo brain MISAR-seq dataset: slice 1, 1,263 spots; slice 2, 1,353 spots; slice 3, 1,777 spots; slice 4, 2,183 spots; slice 5, 1,949 spots; slice 6, 1,939 spots; slice 7, 2,129 spots; slice 8, 2,248 spots;
(3) 2 slices from P22 mouse brain Spatial-CUT&Tag-RNA-seq dataset: slice 1, 9,548 spots; slice 2, 9,370 spots;
(4) 1 slices from P22 mouse brain Spatial-ATAC-RNA-seq dataset: 9,215 spots;
(5) 1 slices from EAE mouse brain Spatial-Mux-seq data dataset: 9,686 spots;
(6) 1 slices from human hippocampus Spatial-ATAC-RNA-seq dataset: 2,500 spots;

(7) E11.5, E12.5, and E13.5 mouse embryo scATAC-seq data: scATAC_Heart, 20,108 cells; scATAC_E13_5_Gonad-male, 20,005 cells; scATAC_ForeBrain, 20,001 cells; scATAC_E13_5_Gonad-female, 20,000 cells; scATAC_Eye, 19,685 cells; scATAC_E12_5_Gonad_male, 18,325 cells; scATAC_SpinalCord, 15,803 cells; scATAC_13_5_Lung, 14,874 cells; scATAC_Kidney, 14,378 cells; scATAC_HindBrain, 13,836 cells; scATAC_Liver, 12,729 cells; scATAC_E12_5_Gonad_female, 12,355 cells; scATAC_Intestine, 12,304 cells; scATAC_Pancreas, 10,316 cells; scATAC_Spleen, 9,296 cells; scATAC_ForeLimb, 7,922 cells; scATAC_E11_5_Gonad-female, 7,695 cells; scATAC_E11_5_Gonad-male, 6,915 cells; scATAC_GermLayer, 6,755 cells; scATAC_Stomach, 5,506 cells; scATAC_12_5_Lung, 4,736 cells; scATAC_MidBrain, 4,674 cells;
(8) E12.5, E13.5, and E14.5 embryonic mouse cerebellum snATAC-seq data: 17,097 cells
(9) E17.5 embryonic mouse heart scATAC-seq data: 2,903 cells
E12.5, E13.5, and E15.5 mouse embryo snATAC-seq data: 1,655 cells;
(10) Three samples of E18 mouse embryo brain snATAC-seq data from 10X Genomics: atac_v1_E18_brain_cryo_5k, 4,747 cells; atac_v1_E18_brain_fresh_5k, 4,030 cells; atac_v1_E18_brain_flash_5k, 3,598 cells
(11) Adult mouse brain snATAC-seq data: 2,355,842 cells
(12) Human brain scATAC-seq data: 70,631 cells
(13) Mouse embryo scRNA-seq data: 2,058,652 cells.
We comprehensively collected scATAC-seq datasets from mouse embryos spanning seven developmental stages and sixteen tissues, providing sufficient coverage to construct a mouse embryo scATAC-seq atlas. We also collected published adult mouse brain scATAC-seq datasets covering 117 brain regions, as well as human brain scATAC-seq datasets spanning 42 regions, enabling the construction of adult mouse and human brain atlases. In addition, we analyzed fourteen spatial epigenomics slices generated using four technologies, five tissue types, and spatial resolutions ranging from 20 to 50. Together, these datasets ensured that our algorithm was evaluated using sufficiently diverse and representative data..

Data exclusions

In Fig. 3c, scOpen was excluded from the benchmark of differential analysis results because scOpen denoises normalized peak matrices rather than raw peak matrices. It is not suitable for differential analysis using the same workflow as applied to raw single-cell/bulk data, and therefore cannot be evaluated using JI.

Replication

All attempts at replication were successful. To make sure that the experimental findings are reproducible, we compared the performance of SPEED with 5 denoising methods on 4 simulated and 14 real SE slices from various sequencing technologies.

Randomization

The experiments were not randomized, because we used all collected data for analysis.

Blinding

Blinding was not used. The datasets analyzed in this study do not involve group allocations (e.g., control versus experimental groups). All experiments were computational, and the execution of the evaluated methods does not permit blinding. Data evaluation was carried out using multiple standard quantitative metrics, with the corresponding calculation procedures detailed in the Methods section.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Clinical data
☒	□ Dual use research of concern
☒	□ Plants

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging

Plants

Seed stocks

N/A

Novel plant genotypes

N/A

Authentication

N/A