PNAS RESEARCH ARTICLE | STATISTICS OPEN ACCESS

Bridging unpaired single-cell multimodal data for integrative analyses with SuperMap

Chao Deng[a,b], Xinyi Ma[c], Hui Lu[a,b], Hongyu Zhao[b,d,1], Jingsi Ming[e,1], and Tao Wang[a,b,c,1]

Affiliations are included on p. 11.

Edited by David Donoho, Stanford University, Stanford, CA; received March 5, 2025; accepted January 7, 2026

Current single-cell profiling technologies enable the capture of multiple cellular modalities, providing valuable insights into complex biological systems. While a substantial amount of single-cell multimodal data has been generated and accumulated, most of these datasets are unpaired, characterized by distinct feature spaces and a lack of cell-wise correspondence. The absence of explicit linkages between modalities poses a fundamental challenge for data integration and interpretation. To address this, we introduce SuperMap, a statistical learning method designed for the integrative analyses of unpaired multimodal data. SuperMap directly learns cross-modal mappings from unpaired data to effectively bridge and link different modalities, facilitating a variety of downstream analysis tasks. Comprehensive benchmarking and real-world applications demonstrate the superior performance of SuperMap in enhancing cell-type identification, improving diagonal integration, enabling regulatory analysis, and revealing epigenomic priming events to specify cell differentiation directions for trajectory inference.

cross-modality data integration | gene activity score | regression with unlinked data

Significance

Integrative analyses of single-cell multimodal data hold immense potential for uncovering valuable biological insights. However, the unpaired nature of such data—where explicit linkages between modalities are absent—poses a fundamental challenge. To address this, we introduce a statistical learning method that directly learns cross-modal mappings from unpaired data. These learned mappings effectively bridge and link different modalities, enabling a wide range of downstream integrative analyses. This approach significantly enhances the interpretability of unpaired multimodal data, unlocking its potential for biological discovery.

Single-cell sequencing technologies have emerged as powerful tools across diverse areas of biomedical research, providing unprecedented insights into cellular heterogeneity and function. Recent technological advancements have enabled the profiling of various cellular modalities at single-cell resolution, including chromatin accessibility (scATAC-seq) (1, 2), DNA methylation (snmC-seq) (3), RNA expression (scRNA-seq) (4, 5), and protein abundance (SCoPE-MS) (6). Integrative analyses of these multimodal data are increasingly important for achieving a comprehensive view of cellular systems and uncovering cross-modal regulatory mechanisms contributing to disease and development.

Several single-cell joint profiling protocols have been developed to simultaneously measure multiple modalities within the same cell (7–10), generating paired multimodal data. Notable examples include 10X Genomics Multiome platform and SHARE-seq (8), which concurrently measure gene expression and chromatin accessibility in a single cell. While multimodal profiling holds great promise, paired data remain relatively scarce and often criticized for their noise, low-throughput, and high cost, which limit their broad applicability (11). Consequently, in most scenarios, single-modality profiling techniques are employed to measure different modalities in biologically similar samples, resulting in the extensive generation and accumulation of unpaired multimodal data. These unpaired data are more accessible and widely available across various species, tissue types, developmental stages, and disease contexts (12, 13). This has driven growing interest in their integrative analyses for various biological applications.

Integrative analyses on unpaired data offer significant advantages but present substantial computational challenges. Since unpaired data are collected in different feature spaces (e.g., chromatin regions in scATAC-seq versus genes in scRNA-seq) and lack cell-wise correspondence, the absence of explicit linkages (or anchors) between modalities poses a critical impediment to downstream biological analyses. For instance, regulatory analysis, which associates features across modalities to uncover regulatory interactions, become infeasible due to the lack of cell-wise pairing (14, 15). The same is true for diagonal integration, which aims to integrate unpaired multimodal data with distinct feature spaces into a unified latent space (16, 17). Existing diagonal integration methods use biological prior knowledge or additional paired multimodal data to establish linkages. For example, popular methods such as Seurat (18), BindSC (19), and scJoint (20) calculate the gene activity score from ATAC data—based on genomic coordinates—to serve as the linkage for aligning RNA and ATAC modalities. Similarly, GLUE (21) employs a prior knowledge-based guidance graph that leverages

Author contributions: C.D., H.Z., J.M., and T.W. designed research; C.D., H.L., H.Z., J.M., and T.W. performed research; C.D., X.M., H.L., and T.W. contributed new reagents/analytic tools; C.D. and X.M. analyzed data; and C.D., H.Z., J.M., and T.W. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: hongyu.zhao@yale.edu, jsming@fem.ecnu.edu.cn, or neowangtao@sjtu.edu.cn.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2505182123/-/DCSupplemental.

Published February 6, 2026.

genomic proximity information to serve as cross-modal linkages. However, such linkages derived from biological priors, rather than being adaptively learned from the data itself, are inherently inaccurate and may introduce biases into the integration results. Other methods, such as Coupled NMF (22), SeuratV5 (23), and StapMap (24), adopt a "mosaic integration" approach by incorporating paired multimodal data from similar biological contexts to serve as linkages. Nevertheless, obtaining suitable paired data is often prohibitively difficult and expensive, limiting the applicability of these methods in real-world scenarios. Consequently, constructing effective cross-modal linkages is crucial and remains a major bottleneck in the integrative analysis of unpaired multimodal data.

To address this, we propose **SuperMap**, a statistical learning method for the integrative analyses of unpaired multimodal data. SuperMap directly learns cross-modal feature mappings from unpaired data. Compared with existing methods for learning such mappings, which require paired data for training and are therefore inapplicable to unpaired settings (25, 26), our framework represents a substantive methodological advancement. The cross-modal mappings learned by SuperMap serve as effective linkages to bridge unpaired data, facilitating a variety of downstream analyses. Specifically, the learned mappings enable us to impute the missing-modality profiles in unpaired datasets, thereby enhancing cell-type identification and enabling regulatory analysis. Additionally, we develop a diagonal integration scheme that utilizes these mappings to construct high-quality alignment anchors, thereby improving integration performance. Furthermore, we apply SuperMap to datasets with continuous cell differentiation, where it reveals epigenomic priming events during the cell differentiation process, through which we identify a population of EOMES+ neuron progenitor cells. Systematic benchmarks and real-world applications demonstrate SuperMap's superior accuracy, robustness, and capacity to yield meaningful biological insights.

Results

The SuperMap Framework. SuperMap is a statistical learning method designed for the integrative analyses of unpaired single-cell multimodal measurements. The input multimodal datasets, which originate from different but biologically similar samples, are each generated through distinct sequencing techniques (e.g., scRNA-seq, scATAC-seq, and snmC-seq), resulting in a lack of cell-wise and feature-wise correspondences (Fig. 1*A*). SuperMap aims to learn cross-modal feature mappings directly from unpaired data. The learned mappings effectively bridge different modalities, filling inherent data gaps and facilitating downstream integrative analyses. Notably, SuperMap operates without the need for suitable paired data, which is often unavailable and costly, thus enhancing its practicality in real-world applications.

The fundamental concept behind SuperMap involves matching the joint distributions of multimodal data to learn cross-modal mappings from unpaired observations. Here we take unpaired scRNA-seq and scATAC-seq data integration as an example for illustration (Fig. 1*B*), though SuperMap is broadly applicable beyond this modality pair. Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_s})^T \in \mathbb{R}^{n_s \times p}$ denote the normalized scRNA-seq expression data and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_t})^T \in \mathbb{R}^{n_t \times q}$ denote the normalized scATAC-seq accessibility data, where each row corresponds to a cell and each column to a feature. Here, n_s and n_t are the numbers of cells in the two datasets, and p and q are the numbers of features. The two datasets have distinct feature sets (genes and chromatin regions, respectively) and lack cell-wise correspondence. Given that both datasets are derived from similar biological systems, they are expected to share an underlying biological structure related to the cell population, such as similar cell-type compositions and cellular state distributions. Let $\mathbf{Y}^* \in \mathbb{R}^{n_t \times p}$ denote the unmeasured RNA profiles for cells in $\mathbf{X}$. Since $\mathbf{Y}$ and $\mathbf{Y}^*$ describe biological systems with similar structure within the same modality, we assume they follow analogous patterns or joint distributions:

$$\mathbf{y} \overset{d}{=} \mathbf{y}^*, \qquad [1]$$

where $\overset{d}{=}$ signifies equality in joint distribution. Furthermore, it is well established that different modalities are not independent, but are linked through cross-modal regulations (27, 28), such as chromatin accessibility influencing transcription. This justifies modeling their association via a regression:

$$\mathbf{y}^* = \mathbf{B}\mathbf{x} + \boldsymbol{\epsilon}, \qquad [2]$$

where $\mathbf{B} \in \mathbb{R}^{p \times q}$ is the coefficient matrix, and $\boldsymbol{\epsilon} \in \mathbb{R}^p$ is the error term. Combining [**1** and **2**], we proposed a statistical learning model for unpaired $\mathbf{y}$ and $\mathbf{x}$, formulated as:

$$\mathbf{y} \overset{d}{=} \mathbf{B}\mathbf{x} + \boldsymbol{\epsilon}.$$

Given the high dimensionality of $\mathbf{y}$ and $\mathbf{x}$, we fitted the model by considering the marginal distributions and pairwise interactions of variables, while incorporating prior knowledge about the regulatory potential of the two modalities to regularize or constrain the model. Furthermore, we designed an efficient and robust optimization algorithm based on the Alternating Direction Method of Multipliers (ADMM) framework (29). Detailed descriptions of the model and algorithm are provided in *Materials and Methods*.

SuperMap can facilitate extensive downstream integrative analyses of unpaired multimodal data, as illustrated in Fig. 1*C*. These include: 1) enhancing cell-type identification by imputing missing-modality profiles through learned mappings—for example, predicting marker genes expression patterns in ATAC cells where such information is absent yet crucial for cell annotation; 2) providing high-quality alignment anchors for diagonal integration, thereby improving modality alignment; 3) enabling regulatory analysis on unpaired data by estimating cross-modal regulatory strengths; and 4) revealing epigenomic priming events that specify cell differentiation directions along the trajectory. Collectively, these capabilities demonstrate the power of SuperMap in analyzing and interpreting unpaired multimodal data, unlocking its full potential for biological discovery.

SuperMap Accurately Learns Cross-Modal Mappings and Imputes Missing-Modality Profiles in Unpaired Data. We evaluated the efficacy of SuperMap in learning cross-modal mappings by assessing its performance in imputing missing-modality profiles (*Materials and Methods*). We employed datasets generated through simultaneous multimodal profiling techniques, which provide the cell pairing information serving as a gold-standard for assessment. In order to assess the generalization capabilities of SuperMap, we used five publicly available datasets derived from various tissues, species, and protocols, which have been extensively used for benchmarking purposes in previous studies (14, 21, 23, 30, 31). These datasets include the 10X Multiome

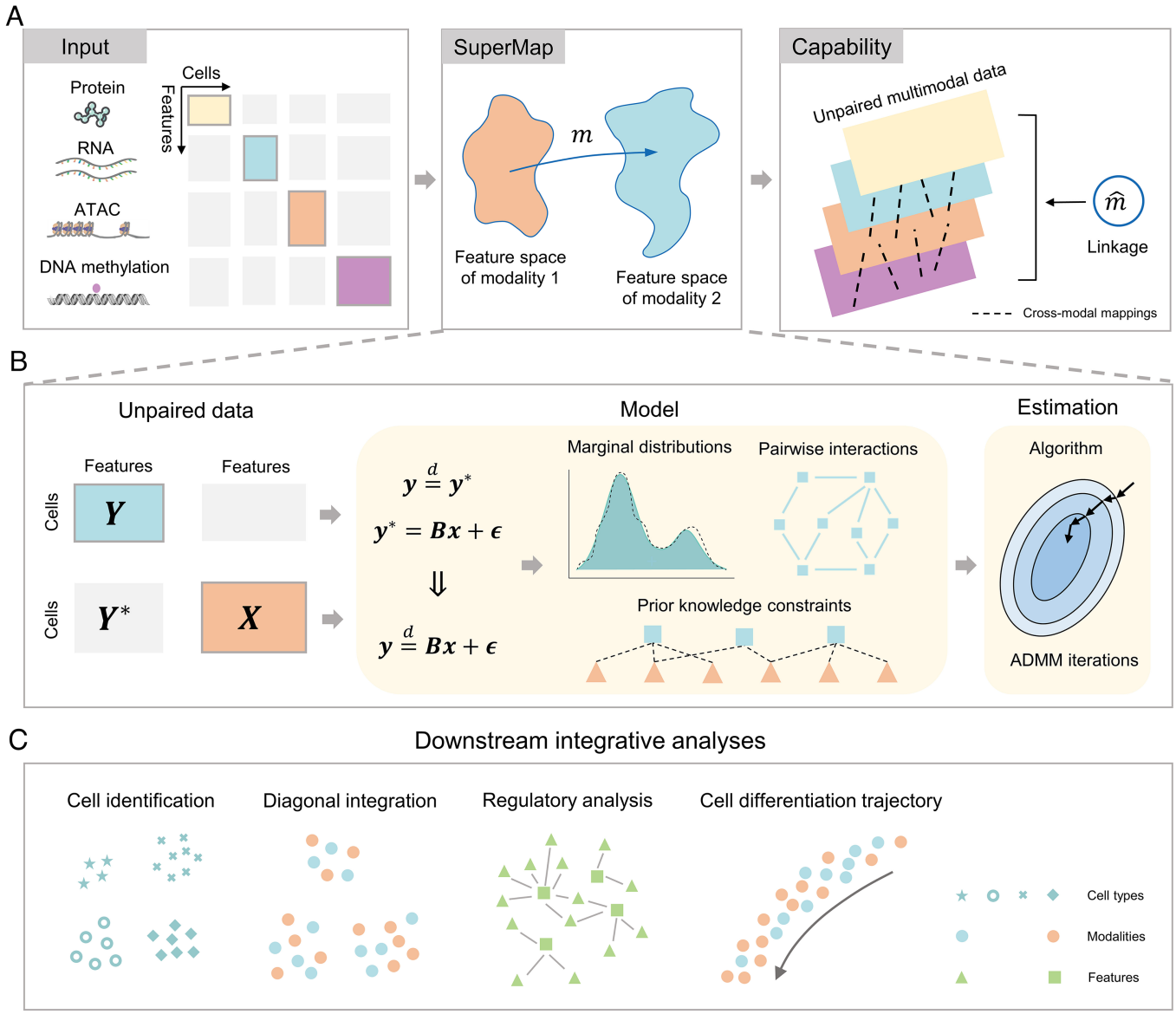

Fig. 1. Overview of SuperMap. (*A*) The input for SuperMap consists of unpaired single-cell multimodal measurements derived from biologically similar samples. SuperMap learns cross-modal feature mappings, which serve as linkages to bridge unpaired data. (*B*) Schematic representation of the SuperMap model. Given unpaired multimodal data (**Y**) and (**X**), a regression-like framework is proposed, formulated as $\mathbf{y} \overset{d}{=} \mathbf{Bx} + \epsilon$. The model is fitted by considering the marginal distributions and pairwise interactions of variables with prior knowledge constraints. An ADMM algorithm is designed for optimization. (*C*) Downstream integrative analyses enabled by SuperMap.

dataset of human peripheral blood mononuclear cells (10X Multiome PBMC), the 10X Multiome dataset of human bone marrow mononuclear cells (10X Multiome BMMC) (32), the SHARE-seq dataset of mouse skin (8), the scNMT-seq dataset of mouse embryos (33), and the ASAP-seq dataset of PBMC (34) (Fig. 2*A* and *SI Appendix*, Fig. S1*A*). Detailed information on the benchmarking datasets is provided in *SI Appendix*, Text. Of note, in our analysis, we intentionally disregarded the inherent pairing between modalities, treating them as unpaired to simulate real-world scenarios, with the cell pairing information used solely for performance evaluation.

We first evaluated the performance of SuperMap in imputing RNA expression from ATAC accessibility. We benchmarked SuperMap against existing gene activity score calculation methods, including Signac (35), MAESTRO (36), and UnpairReg (30), on 10X Multiome PBMC, 10X Multiome BMMC, and SHARE-seq mouse skin datasets. We employed a paired regression model that utilizes paired information as the gold standard reference for performance comparison. To evaluate the imputation accuracy, we used three metrics: Pearson's correlation, Spearman's correlation, and mean squared error (MSE). Across all three datasets, SuperMap demonstrated performance metrics closely aligned with those of paired regression and significantly outperformed existing methods (Fig. 2*B*). For the SHARE-seq mouse skin dataset, the shallower sequencing depth resulted in lower imputation accuracy for all methods; however, the improvement achieved by SuperMap remained evident. Moreover, we identified a set of marker genes for each cell type in the PBMC dataset (*SI Appendix*, Table S1), following Seurat pipeline. We found that all methods generally performed better on the marker genes than on nonmarker genes (Fig. 2*C*), indicating that marker-related signals are reliably captured. Notably, SuperMap

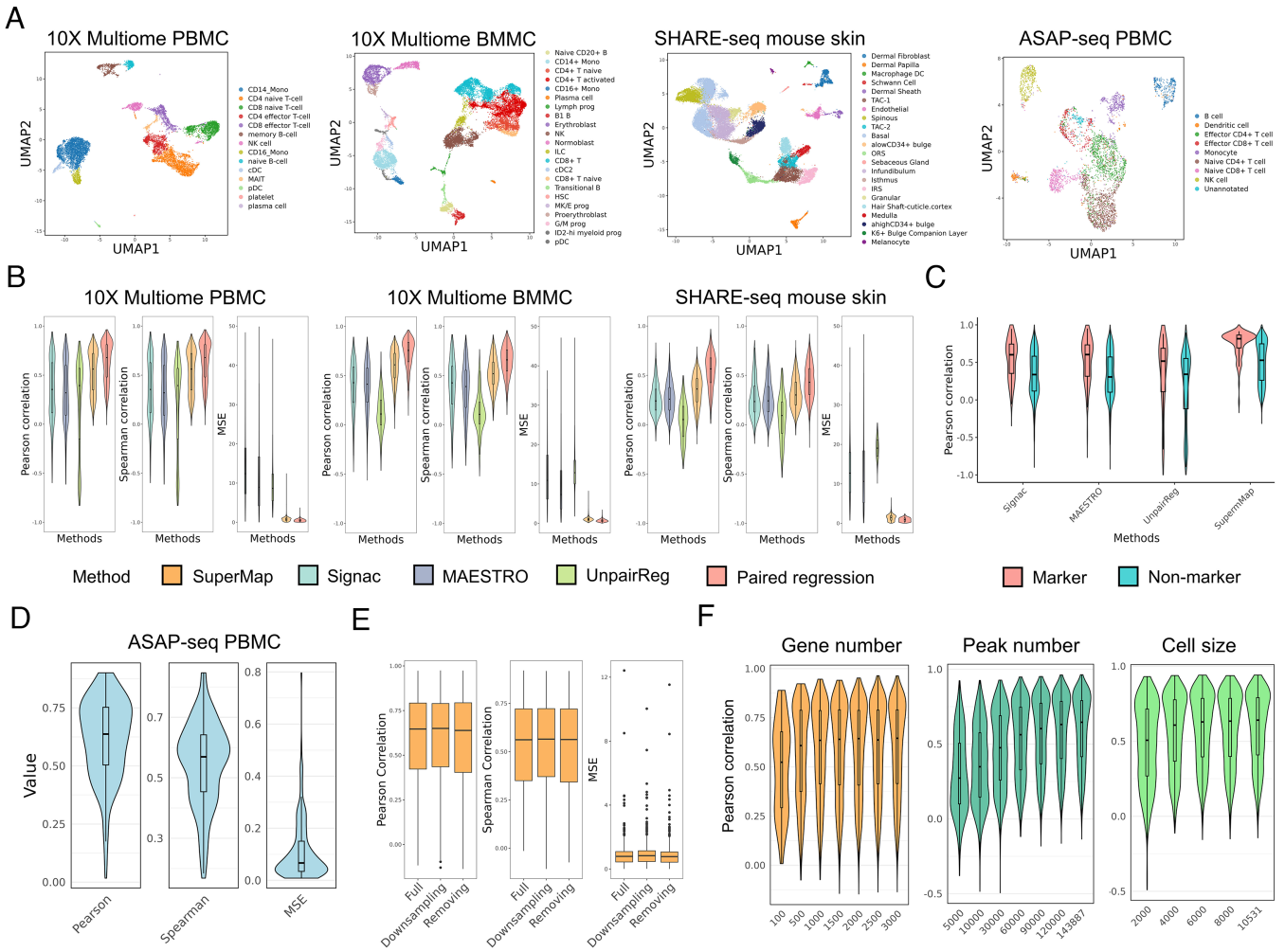

Fig. 2. Benchmarking the performance of SuperMap in learning cross-modal mappings. (*A*) UMAP visualization of benchmark datasets (10X Multiome PBMC, 10X Multiome BMMC, SHARE-seq mouse skin and ASAP-seq PBMC), colored by cell types. (*B*) Evaluation of imputation accuracy for predicting RNA expression from ATAC accessibility across different methods on benchmarking datasets. Each dot in the violin plot corresponds to one gene. (*C*) Comparison of the imputation performance of different methods on marker versus nonmarker genes. (*D*) Imputation accuracy of SuperMap in predicting protein abundance from chromatin accessibility on ASAP-seq PBMC data. (*E*) Imputation accuracy achieved by SuperMap on imbalanced datasets. (*F*) Imputation accuracy of SuperMap under varying modality feature dimensions and cell sizes.

exhibited superior imputation performance compared with the other methods, highlighting its potential to enhance cell-type identification. For many marker genes, SuperMap successfully recovered the expected expression patterns where the other methods failed (*SI Appendix*, Fig. S2). For several highly specific markers, the baseline methods generally performed reasonably well; nevertheless, SuperMap also exhibited strong cell-type specificity. Collectively, these results demonstrate that SuperMap accurately learns RNA–ATAC mappings from unpaired data and enhances cell-type identification.

While RNA–ATAC integration is one of the most common and widely applied scenarios in biomedical research, the underlying concept of SuperMap—matching joint distributions—is broadly applicable to other modality pairs. To assess its generalizability, we applied SuperMap to integrate protein measurements with chromatin accessibility data using the ASAP-seq PBMC dataset, which contains antibody-derived tag measurements for 227 proteins along with genome-wide chromatin accessibility profiles. We treated protein levels as the response variable $\mathbf{y}$ and chromatin accessibility as the predictor variable $\mathbf{x}$. We evaluated SuperMap's performance in imputation accuracy. As shown in Fig. 2*D*, SuperMap achieved satisfactory performance, with accuracy comparable to that observed in the RNA–ATAC integration tasks. To our knowledge, no methods currently exist for directly imputing protein abundance from ATAC data in unpaired context, so we did not include baseline comparisons for this. Furthermore, since many of the measured proteins are classic cell surface markers (e.g., CD4, CD14, and CXCR5), we visualized the imputed values of these markers. We found that SuperMap accurately captured expected expression patterns (*SI Appendix*, Fig. S1*C*). These findings demonstrate that SuperMap can be effectively applied to the protein–ATAC modality pair to learn their cross-modal mappings.

To further investigate SuperMap's performance in predicting RNA expression from DNA methylation, we applied it to the scNMT-seq dataset of mouse embryos. Unlike chromatin accessibility, DNA methylation levels typically exhibit an inverse relationship with gene expression (37, 38). Therefore, SuperMap utilized demethylation data for this analysis. The mean demethylation rate (MDR), defined as the proportion of unmethylated CpGs to the total CpGs in the promoter region (extending 2 kb upstream and downstream of the transcription start sites) (39),

served as the baseline method. We evaluated the imputation accuracy using Pearson's correlation between the predicted RNA expression and the ground truth for each gene. As shown in *SI Appendix*, Fig. S1*B*, SuperMap exhibited more accurate imputation for the majority of genes, demonstrating its superiority.

Robustness of SuperMap to Data Imbalance and Varying Data Conditions. In the above evaluations, cell types were consistent and cell-type proportions were balanced between two modalities. However, this balance is not always guaranteed. In practice, even when two modalities are biologically matched, technical factors like sampling variation or sequencing bias can introduce mild discrepancies in cell-type proportions. Since SuperMap learns cross-modal mappings adaptively from the data, its sensitivity to data imbalance warrants further investigation. Therefore, we conducted data corruption studies to evaluate how SuperMap performs when cell types are unmatched or cell-type proportions are imbalanced. We constructed two additional datasets based on the 10X Multiome PBMC dataset: 1) downsampling the three most prevalent cell types (CD14_Mono, CD4 naive T-cell, and CD8 naive T-cell) in the scATAC-seq data by removing 50% of the cells while keeping the scRNA-seq data unchanged; 2) removing the three least common cell types (pDC, plasma cell, and platelet) from the scATAC-seq data while maintaining all the scRNA-seq data. To aid interpretation, we included UMAP visualizations of the corrupted ATAC data, which clearly show the absence of the removed cells, as indicated by the red arrows (*SI Appendix*, Fig. S3*A*). These artificially generated datasets were used to fit the model. Despite the induced imbalance, the imputation accuracy of SuperMap remains comparable to that achieved from the original complete data (Fig. 2*E*), demonstrating SuperMap's robustness to cell-type proportion imbalance. We also evaluated the cell-type-specific imputation accuracy. As shown in *SI Appendix*, Fig. S3*B*, SuperMap consistently maintained stable performance across all cell types. The key to this robustness lies in that SuperMap learns cross-modal mappings by fitting the marginal distributions and pairwise interactions of features. Under above mild data imbalance conditions, while the marginal distributions exhibited slight shifts, the covariance structure of the data is largely preserved which provides critical information for accurate mapping estimation (*SI Appendix*, Fig. S4 *A* and *B*). Additional experiments show that as severe imbalance increasingly distorts the data distribution, imputation accuracy progressively declines (*SI Appendix*, Fig. S4*C*); see *SI Appendix*, Text for details.

We also investigated how modality-specific feature dimensions and cell size affect SuperMap's performance. The results show that when the number of response features is small, imputation accuracy decreases markedly (Fig. 2*F* and *SI Appendix*, Fig. S5*A*), indicating that an adequate number of features is required to reliably learn cross-modal mappings. A similar trend is observed for predictor modality (Fig. 2*F* and *SI Appendix*, Fig. S5*B*). Regarding cell size, SuperMap's imputation accuracy significantly declines when the number of cells used for model fitting is small (Fig. 2*F* and *SI Appendix*, Fig. S5*C*). Detailed explanations are provided in *SI Appendix*, Text. Finally, we compared the runtime performance of SuperMap and the baseline methods across varying data sizes. SuperMap's runtime scales linearly with sample size (*SI Appendix*, Fig. S5*D*); further details can be found in *SI Appendix*, Text.

SuperMap Facilitates Diagonal Integration by Providing Effective Alignment Anchors. While several methods have been developed for diagonal integration, they primarily focus on aligning modalities by removing modality-specific effects. Rather than proposing yet another alignment algorithm, we focus on constructing high-quality anchors—the major bottleneck in this field. To operationalize this idea, we designed a diagonal integration scheme (*Materials and Methods*), which offers a perspective and conceptual innovation for addressing the diagonal integration problem.

We conducted a systematic comparative evaluation of SuperMap against several state-of-the-art integration methods, including Seurat, MAESTRO, BindSC, GLUE, and scJoint, all of which have demonstrated strong performance in benchmarking studies (31). The benchmarking datasets employed in the preceding section were utilized for this evaluation. The UMAP visualization of the integrated embeddings showed that SuperMap effectively aligned the two modalities, with cell groupings closely corresponding to cell-type labels (Fig. 3*A* and *SI Appendix*, Figs. S6 *A* and *B* and S7*A*).

To quantitatively assess integration performance, we computed several evaluation metrics across three key aspects: cell alignment error, label transfer accuracy, and biological conservation. Cell alignment error was quantified using the FOSCTTM metric (21, 40), with lower values indicating better modality alignment. Label transfer accuracy was evaluated using four commonly adopted metrics: cell annotation accuracy (ACC), adjusted rand index (ARI), normalized mutual information (NMI), and average F1 score (F1) (41). To assess biological conservation—specifically, how effectively an integration method preserves biological signals—we employed the KNN purity score (KNN_score) (23) and mean average precision (MAP) (21) metrics. These six metrics range from 0 to 1, with higher values indicating better performance (see *SI Appendix*, Text for details). As shown in Fig. 3 *B* and *C*, on the PBMC and BMMC benchmarking datasets, SuperMap consistently outperformed competing methods across all evaluation metrics. However, on the SHARE-seq mouse-skin dataset, SuperMap achieved the second-best performance, likely due to the dataset's lower per-cell sequencing depth (Fig. 3*D*). Such low-depth data make it difficult for SuperMap to perform reliable imputation at single-cell resolution, leading to degraded integration performance. Additionally, the consistently superior performance of SuperMap over Seurat demonstrates that, when provided with reliable alignment anchors, even previously less effective integration methods can achieve satisfactory performance (Fig. 3 *B* and *C*).

Furthermore, as an add-on module, SuperMap can also be incorporated into other integration methods by providing effective alignment anchors. We combine it with BindSC, another integration approach that, like Seurat, uses the gene activity score as linkage. SuperMap augmented BindSC showed substantially improved performance compared to its original implementation (Fig. 3*E*). This further highlights the effectiveness of SuperMap-derived anchors and underscores the crucial role of anchor quality in diagonal integration.

As described in *Materials and Methods*, we performed imputation for diagonal integration at single-cell resolution; however, we used meta-cell data for model fitting, in which the choice of meta-cell graining level is an important factor to investigate. The graining level refers to the degree of size reduction from single-cell to meta-cell data (i.e., single-cell number/meta-cell number). A graining level of 50 corresponds to a 50-fold reduction. Following previous studies (42), we set the graining level to around 60 to 70. To further examine its effect, we systematically evaluated the impact of graining level on model fitting. Using the 10X Multiome PBMC dataset, we constructed meta-cell data with varying graining levels (from 25 to 105), fitted the model for each, and evaluated downstream integration quality. The results

show that SuperMap performs very consistently across varying graining levels (Fig. 3*F* and *SI Appendix*, Fig. S7 *B* and *C*). This is expected, as meta-cell data with different graining levels contain the same information about feature mappings, allowing SuperMap to learn these mappings and yield similar downstream results. These findings suggest that SuperMap is robust to the choice of graining level, and values within a reasonable range (25 to 105) are feasible. The specific choice can be adjusted according to data characteristics—for example, a higher graining level for large or low-depth datasets.

SuperMap Enables Regulatory Analysis on Unpaired Multimodal Data. In the study of development and disease, dissecting regulatory interactions is fundamental for understanding key biological processes. Traditional approaches for regulatory analysis predominantly rely on paired data (14, 15), which is often unavailable and costly to obtain. In contrast, benefiting from its ability to learn cross-modal mappings, SuperMap enables regulatory analysis using unpaired data (*Materials and Methods*). To demonstrate this, we applied SuperMap to the 10X Multiome PBMC dataset, treating it as unpaired scRNA-seq and scATAC-seq data. SuperMap successfully identified several significant peak-gene linkages. As shown in Fig. 4*A*, the linked peak and gene exhibited strong concordant patterns in chromatin accessibility and gene expression levels. Furthermore, a negative correlation was observed between the regulatory score and genomic distance of the peak-gene pairs, as expected, with regulatory elements predominantly enriched in proximity to the

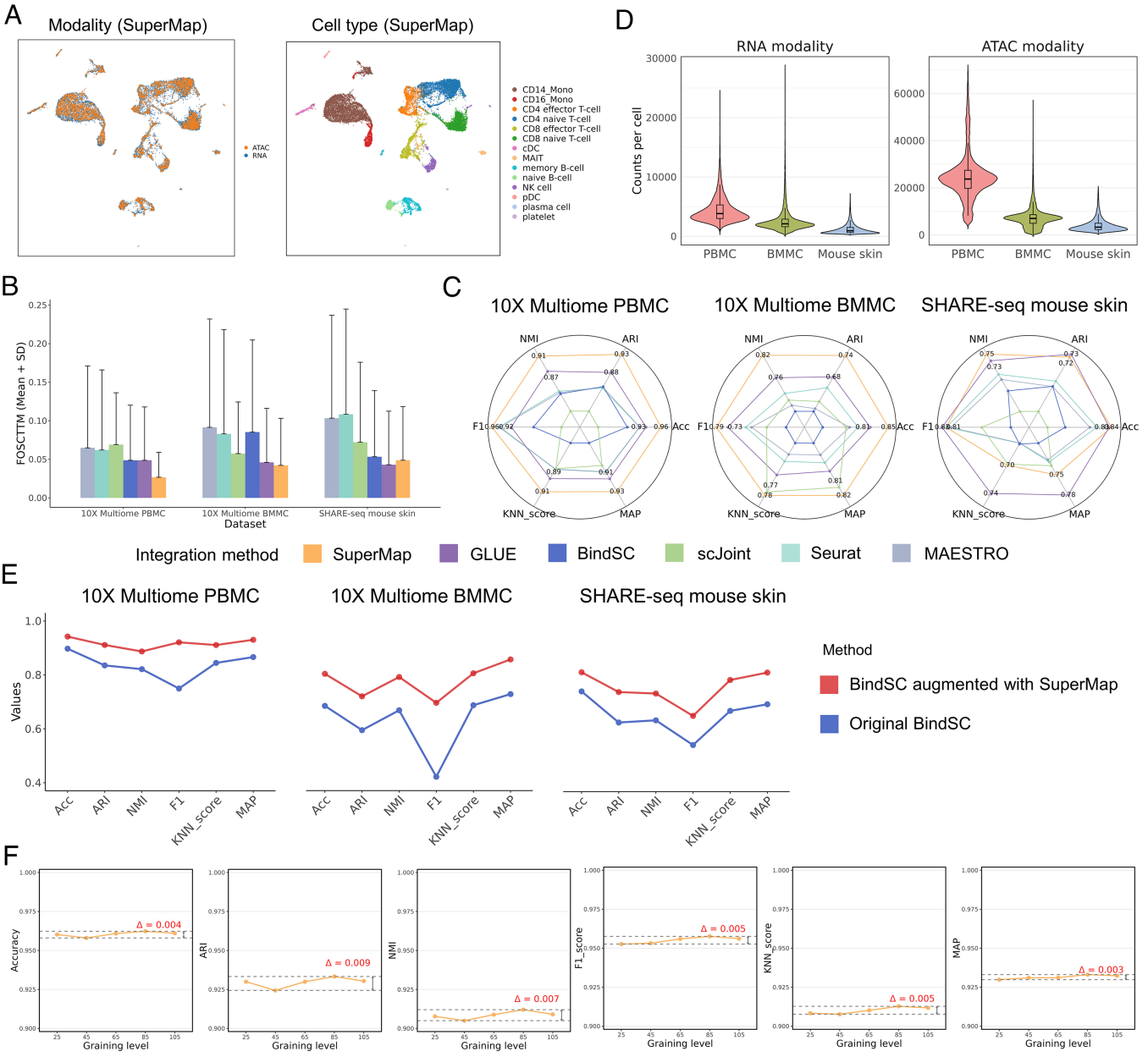

Fig. 3. Evaluation of SuperMap's performance on diagonal integration. (*A*) UMAP visualization of integrated embeddings generated by SuperMap on the PBMC dataset. (*B*) Cell alignment error (quantified by FOSCTTM) of SuperMap and competing methods across benchmarking datasets. (*C*) Quantitative evaluation of label transfer accuracy and biological conservation for SuperMap and competing methods across benchmarking datasets. (*D*) Per-cell sequencing depth across benchmark datasets. (*E*) Comparison of integration performance between original BindSC and SuperMap augmented BindSC. (*F*) Quantitative evaluation of integration performance across different meta-cell graining levels.

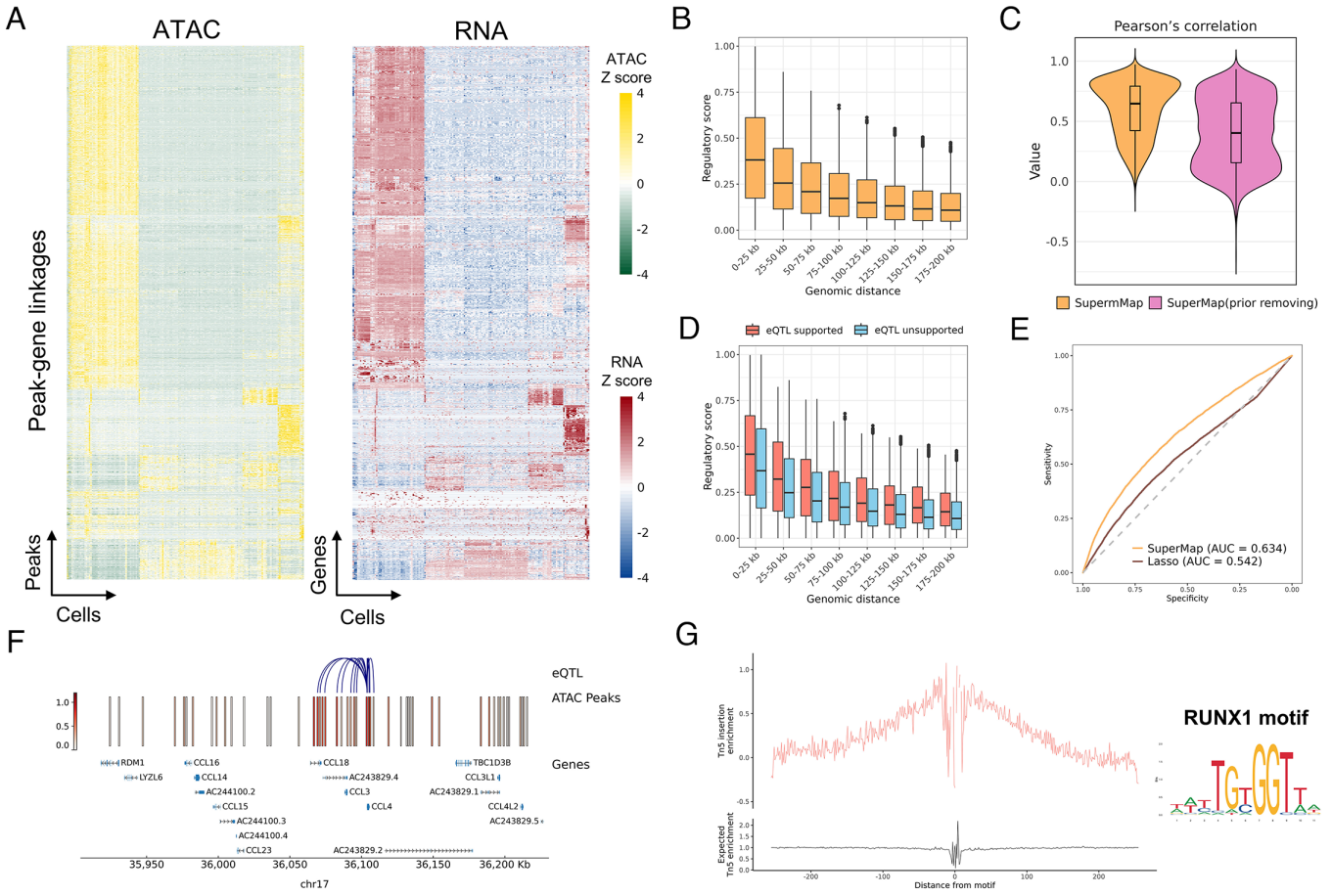

Fig. 4. Regulatory analysis with SuperMap. (*A*) Heatmaps of chromatin accessibility (*Left*) and measured gene expression (*Right*) for significant peak-gene linkages (regulatory score >0.5). Each row represents a pair consisting of one peak and one linked gene, while each column represents a cell. (*B*) SuperMap-estimated regulatory scores for peak-gene pairs across varying genomic distance ranges. (*C*) Comparison of SuperMap's imputation accuracy with and without incorporating prior knowledge. (*D*) SuperMap-estimated regulatory scores for peak-gene pairs across varying genomic distance ranges, categorized by the presence or absence of eQTL support. (*E*) ROC curve for eQTL interactions prediction by varying regulatory scores using SuperMap and Lasso regression. (*F*) Genome track centered around the CCL4 gene. The x-axis represents genomic coordinates. Blue arcs at the top of the panel indicate eQTL-gene regulatory associations involving CCL4. Each arc connects an eQTL locus (the distal end of the arc) to the TSS of CCL4 (the common anchor point). The color intensity in the ATAC peaks track reflects the magnitude of the estimated regulatory scores. (*G*) TF footprints for RUNX1 motif.

transcription start site (0 to 25 kb) (Fig. 4*B*). However, since our model incorporates genomic proximity priors, it is critical to assess its impact on regulatory analysis. To this end, we removed the prior from SuperMap and reevaluated its performance. We observe a decline in imputation accuracy when the prior was omitted (Fig. 4*C* and *SI Appendix*, Fig. S8*A*), indicating that the prior plays a beneficial role in model. Notably, the negative correlation between regulatory scores and genomic distance persist (*SI Appendix*, Fig. S8*B*).

To further validate the predicted cis-regulatory interactions, we incorporated external validation using expression quantitative trait loci (eQTL) data derived from whole blood, obtained from the GTEx database (43). We assessed the concordance between the SuperMap regulatory scores calculated from the PBMC dataset and the eQTLs, considering a peak-gene pair as eQTL-supported if 1) the peak overlapped with an eQTL locus and 2) the eQTL locus exhibited a significant association with the corresponding gene. Across all genomic distances, regulatory scores for eQTL-supported pairs were consistently higher compared to those without such support (Fig. 4*D*). When considering eQTLs as a benchmark, SuperMap achieved a higher area under the curve (AUC) compared to Lasso regression, indicating that the regulatory score accurately reflects cis-regulatory interactions (Fig. 4*E*). For example, CCL4 (C–C motif chemokine ligand 4), encoded by the CCL4 gene (located at chr17:36103827-36105621), is a well-known chemokine that functions as a potent chemoattractant for various immune cells, including NK cells, macrophages, and T cells (44), and plays a critical role in both inflammatory and immune responses (45, 46). The eQTLs significantly associated with CCL4 expression are predominantly located within the 0 to 25 kb upstream region of the gene body (Fig. 4*F*). Notably, peaks overlapping these eQTLs had relatively higher regulatory scores, highlighting the effectiveness of SuperMap in identifying cis-regulatory interactions. Furthermore, we conducted motif enrichment analysis based on significant peaks linked to CCL4 using ArchR (47) with motif database CISBP(v2) (48). The transcription factor (TF) RUNX1 motif emerged as the most highly enriched, suggesting that RUNX1 may regulate CCL4 expression by binding to these cis-regulatory regions. TF footprinting analysis further confirmed the binding of RUNX1 to these regions (Fig. 4*G*).

SuperMap Reveals Epigenomic Priming Events and Delineates Cell Differentiation Direction. To investigate the utility of SuperMap in integrating single-cell multimodal datasets with

continuous cell state changes, we applied it to neurogenesis-related sciRNA-seq3 data (49) and sciATAC-seq3 data (50) derived from the human fetal cerebrum (*SI Appendix*, Fig. S9*A*). The datasets comprise samples from human fetuses aged 89 to 122 d postconception, with gene expression and chromatin accessibility data generated separately.

Using SuperMap, the two modalities were successfully integrated into a unified latent space and the aligned cell groupings were largely consistent with the original annotations (Fig. 5*A*). This integration further facilitated the refinement of cell annotations. For instance, a group of ATAC cells that were originally annotated ambiguously as "Astrocytes/Oligodendrocytes" was partitioned into two distinct populations after integration, aligning respectively with the "Astrocytes" and "Oligodendrocytes" clusters in the RNA data (highlighted by red dashed circles in Fig. 5*A*). However, we also noted some inconsistencies: a subset of ATAC cells annotated as "Astrocytes" aligned with a portion of "Excitatory neurons" in the RNA data (indicated by black arrows in Fig. 5*A*). Further analysis revealed that this subset actively expressed classical radial glial cell markers, including GLI3, HES1, and HOPX (51, 52), suggesting that these cells may represent a population of radial glial cells rather than the originally annotated cell types (Fig. 5*B*).

Next, we constructed the cell differentiation trajectory using Monocle3 (53) based on the integrated embeddings. Our analysis specifically focused on the excitatory neuron lineage, which constitutes the primary structure of the inferred trajectory (Fig. 5*C*). To perform pseudotime ordering, prior knowledge of the differentiation direction is typically required to specify the initial state, e.g., root or progenitor cell. However, such information is often lacking or uncertain in many cases. By integrating unpaired multimodal data, SuperMap demonstrated its capacity to infer differentiation direction by exploring how sequential changes in one modality correlate with those in the other. We first identified a set of differentiation-dependent genes that exhibited significant dynamic changes along the

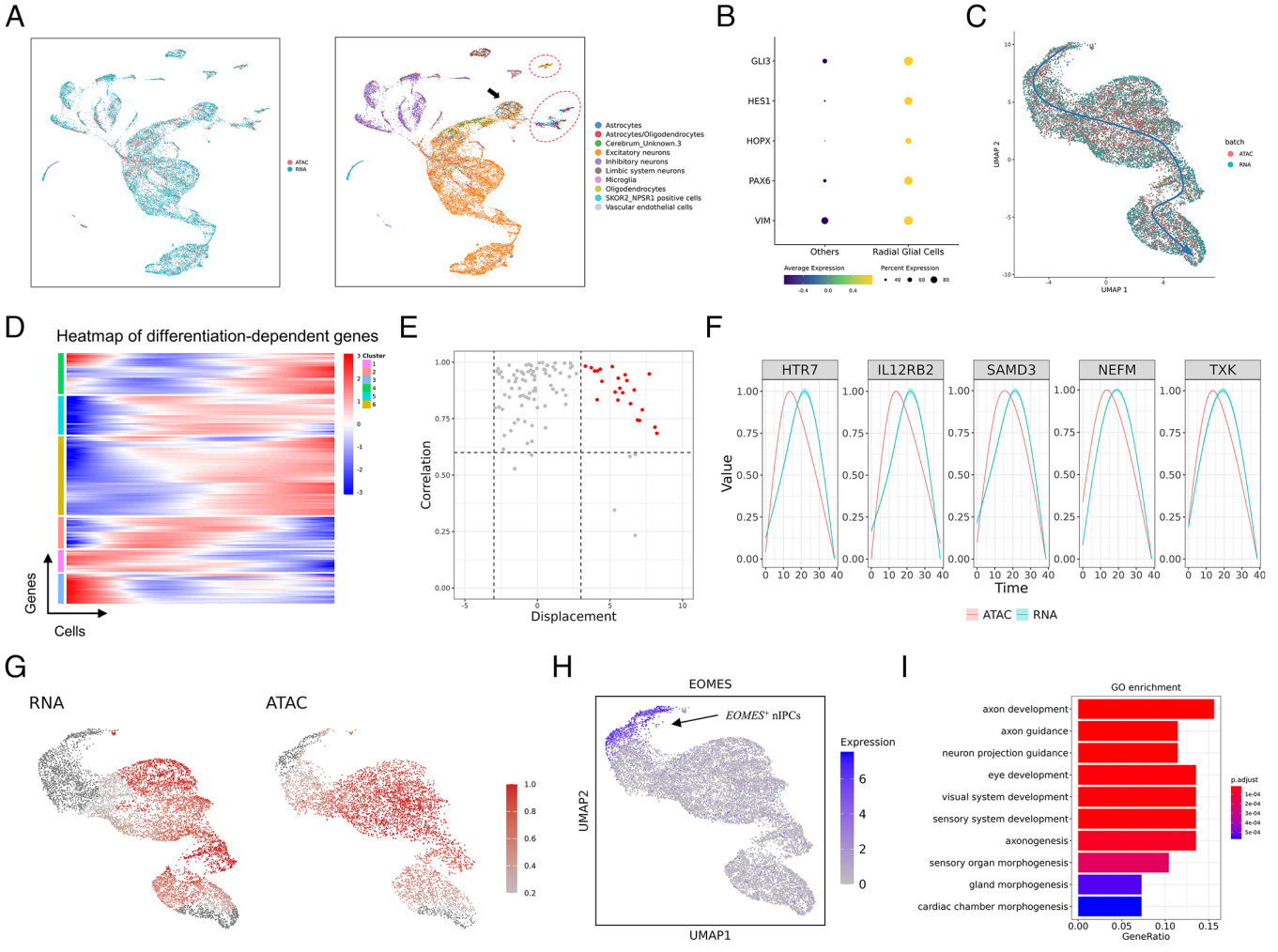

Fig. 5. SuperMap captures epigenomic priming by integrating multimodal datasets of human fetal cerebrum. (*A*) UMAP visualizations of integrated cell embeddings, colored by modalities (*Left*) and original cell annotations (*Right*). (*B*) Dot plots for the average expression levels of marker genes in radial glial cells and compared to other cell types. (*C*) Inferred trajectory of excitatory neurons based on the integrated RNA and ATAC data. (*D*) Heatmap for the expression of differentiation-dependent genes along the inferred trajectory, with genes clustered based on their dynamic patterns. (*E*) Temporal displacement between ATAC-predicted gene expression inferred via SuperMap and experimentally measured RNA expression. A displacement is considered significant if it exceeds 3 pseudotime units and the correlation between the two expression profiles is greater than 0.6. (*F*) Spline-smoothed curves of ATAC-predicted gene expression and measured RNA expression along pseudotime, with 95% confidence bands, for representative genes showing temporal displacement. (*G*) Comparison of RNA expression and the ATAC-predicted gene expression for the HTR7 gene along the trajectory. (*H*) Expression pattern of EOMES in the excitatory neuron lineage. (*I*) Top 10 enriched Gene Ontology biological process terms of actively expressed marker genes in EOMES+ nIPCs. One-sided Fisher's exact test was used with P values adjusted using the Benjamini–Hochberg method.

inferred trajectories (Fig. 5*D*) (*Materials and Methods*). Among these genes, several demonstrated notable temporal displacement between the ATAC-predicted gene expression inferred via SuperMap and experimentally measured RNA expression during differentiation (Fig. 5 *E* and *F*) (*Materials and Methods*). Importantly, the consistency in the direction of these temporal displacements suggests the presence of epigenomic priming (54, 55), where changes in chromatin state precede RNA expression, i.e., chromatin opening precedes RNA production. This implies that along the correct differentiation direction of the trajectory, RNA expression will lag behind the ATAC-predicted gene expression. A representative example is HTR7, a gene encoding the serotonin receptor, which aids in neural circuit development by promoting axon growth and synaptogenesis (56). During early differentiation stages, chromatin opening begins, yet HTR7 RNA expression remains relatively low, even as its accessibility progressively increases toward the terminal stage (Fig. 5*G*). Similarly, in later stages, after chromatin closure begins, there is a delay before the HTR7 RNA levels start to decline. This suggests that the chromatin state acts as a precursor to RNA expression, with epigenetic changes foreshadowing subsequent gene expression events. Furthermore, regulatory analysis conducted using SuperMap revealed that identified regulatory elements were activated prior to RNA expression along the pseudotime trajectory (*SI Appendix*, Fig. S9*B*). These findings highlight SuperMap's ability to reveal epigenomic priming and offer valuable insights into cell differentiation direction.

Empowered by these findings, we identified a population of EOMES+ neuron intermediate progenitor cells (nIPCs) at the initiation of the differentiation trajectory, which were previously annotated as "Cerebrum_Unknown.3" (Fig. 5*H*). Consistent with previous studies (57, 58), EOMES+ nIPCs function as neural stem cells, differentiating into various mature excitatory neurons. To further characterize this cell population, we examined their actively expressed marker genes (*SI Appendix*, Fig. S9*C*). Gene Ontology (GO) enrichment analysis revealed these genes are primarily involved in developmental processes, particularly those related to neuron development, further supporting their functional role (Fig. 5*I*).

Discussion

The extensive generation and accumulation of unpaired single-cell multimodal data present significant opportunities for advancing biological insights, but they also pose challenges in data analysis and interpretation. To address these, we introduce SuperMap, a statistical learning method designed for the integrative analyses of unpaired multimodal data. SuperMap directly learns cross-modal mappings from unpaired data using a regression-like framework. We then propose a penalized learning criterion that focuses on the marginal distributions and pairwise interactions of the variables with prior knowledge constraints. Furthermore, we develop an efficient optimization algorithm within the ADMM framework to solve this problem.

By learning cross-modal mappings, SuperMap effectively bridges and links different modalities, facilitating a variety of downstream biological analyses, including cell-type identification, diagonal integration, regulatory inference, and trajectory analysis. Comprehensive benchmarking and real-world applications demonstrate SuperMap's superior accuracy, robustness, and its ability to uncover meaningful biological insights. Our work substantially enhances the interpretability of unpaired multimodal data, unlocking its potential for biological discovery and demonstrating strong practical utility.

In this paper, when modeling RNA and ATAC modalities, we consistently adopt this RNA∼ATAC regression direction, rather than the reverse (i.e. ATAC∼RNA). This design choice was made based on several considerations. First, regulation between modalities is inherently directional—chromatin accessibility affects its targeted gene expression, not the other way around, and so modeling the reverse direction would conflict with this causality. Second, predicting gene expression from chromatin accessibility is more useful and interpretable for downstream biological analyses. This is because ATAC data is inherently less interpretable—much less is known about the functional roles of genomic regions, particularly noncoding regions, compared to well-characterized genes. Third, ATAC data are typically much higher-dimensional and noisier than RNA data, making them more challenging to predict. To demonstrate this, taking advantage of the paired nature of the 10X Multiome PBMC dataset, we fit paired regression models in both directions—RNA∼ATAC and ATAC∼RNA—and compare their predictive performance. We observed that RNA∼ATAC yields substantially better predictive accuracy (*SI Appendix*, Fig. S10). In contrast, ATAC∼RNA performs poorly, with prediction accuracy close to random, suggesting very limited signal in predicting ATAC profiles from RNA. Together, we conclude that RNA∼ATAC is more appropriate for SuperMap than the reverse.

Materials and Methods

The SuperMap Model. Let $\mathbf{Y} = (\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{n_s})^T \in \mathbb{R}^{n_s \times p}$ represent the normalized data of modality 1, such as the scRNA-seq data, and $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{n_t})^T \in \mathbb{R}^{n_t \times q}$ the normalized data of modality 2, such as the scATAC-seq data. Here, n_s and n_t denote the numbers of cells in the respective datasets, while p and q represent the numbers of features. The matrix $\mathbf{Y}^* \in \mathbb{R}^{n_t \times p}$ denotes the unmeasured data of modality 1 for the cells in $\mathbf{X}$. We assume the existence of a mapping function $m : \mathbb{R}^q \to \mathbb{R}^p$ such that the prediction of $\mathbf{Y}^*$ based on $\mathbf{X}$ is given by $\hat{\mathbf{Y}}^* = m(\mathbf{X})$. As mentioned above, existing methods for learning such mappings require paired data for training and are therefore inapplicable to unpaired settings. To address this, we propose a statistical learning model for unpaired data:

$$\mathbf{y} \overset{d}{=} \mathbf{B}\mathbf{x} + \boldsymbol{\epsilon}, \qquad [3]$$

where $\overset{d}{=}$ signifies equality in distribution, $\mathbf{B} = (B_{ij}) \in \mathbb{R}^{p \times q}$ denotes the coefficient matrix, and $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_p)^T \in \mathbb{R}^p$ is an error term with a mean vector of zeros and a covariance matrix $\boldsymbol{\Sigma} = (\Sigma_{ij})$. This model extends the *unlinked* or *unmatched* regression framework typically applied to univariate variables (59, 60). As discussed in *Results*, this formulation is motivated by two biological observations: 1) the shared underlying biological structure between modalities justifies the distributional equivalence $\mathbf{y} \overset{d}{=} \mathbf{y}^*$; and 2) the regulatory links between modalities justify the regression $\mathbf{y}^* = \mathbf{B}\mathbf{x} + \boldsymbol{\epsilon}$. Combining these two biological motivations, we obtain our working model [**3**].

For a random variable $z \in \mathbb{R}$, its cumulative distribution function is denoted by $\mathbb{F}^z(z_0) = \mathbb{P}(z \le z_0)$. Similarly, for a vector $\mathbf{z} \in \mathbb{R}^p$, $\mathbb{F}^{\mathbf{z}}(\mathbf{z}_0) = \mathbb{P}(\mathbf{z} \le \mathbf{z}_0)$. Model [**3**] can then be equivalently expressed as:

$$\mathbb{F}^{\mathbf{y}} = \mathbb{F}^{\mathbf{B}\mathbf{x}} \star \mathbb{F}^{\boldsymbol{\epsilon}},$$

where $\star$ denotes the convolution operator. Fitting this model is highly challenging due to the multivariate nature of both $\mathbf{y}$ and $\mathbf{x}$, with the high dimensionality adding further complexity. To address this, we focus on the marginal distributions and pairwise interactions of the variables, proposing minimizing the following criterion:

$$\min_{\mathbf{B},\boldsymbol{\Sigma}\succ 0}\sum_{i=1}^{p}\left[\int\left\{\mathbb{F}_{n_s}^{y_i}(y_i)-\left(\mathbb{F}_{n_t}^{\boldsymbol{\beta}_i^T\mathbf{x}}\star\mathbb{F}^{\epsilon_i}\right)(y_i)\right\}^2 d\mathbb{F}_{n_s}^{y_i}(y_i)\right]+\frac{w}{2}\left\|\frac{1}{n_s}\mathbf{Y}^T\mathbf{Y}-\boldsymbol{\Sigma}-\frac{1}{n_t}\mathbf{B}\mathbf{X}^T\mathbf{X}\mathbf{B}^T\right\|_F^2,$$

where the subscripts n_s and n_t indicate the corresponding empirical distributions, $\mathbb{F}^{\epsilon_i}$ is specified later, $\boldsymbol{\beta}_i^T$ denotes the i-th row of $\mathbf{B}$, and $\|\cdot\|_F$ is the Frobenius norm. The weight w is used to balance the contribution of the marginal distributions and pairwise interactions. The restriction to the first two moments of the joint distribution results in a loss of information, except when $\mathbf{y}$ follows a multivariate normal distribution. Nevertheless, in many practical scenarios, this approach is reasonable because these moments capture critical information for downstream analyses. For example, for gene expression profiles, marginal distributions are required to identify marker genes and perform differential expression analysis, while pairwise interactions are necessary for regulatory network inference or coexpression analysis.

To incorporate prior knowledge about the regulatory potential of the two modalities, we consider the regularized criterion:

$$\min_{\mathbf{B},\boldsymbol{\Sigma}\succ 0}\sum_{i=1}^{p}\left[\int\left\{\mathbb{F}_{n_s}^{y_i}(y_i)-\left(\mathbb{F}_{n_t}^{\boldsymbol{\beta}_i^T\mathbf{x}}\star\mathbb{F}^{\epsilon_i}\right)(y_i)\right\}^2 d\mathbb{F}_{n_s}^{y_i}(y_i)\right]+\frac{w}{2}\left\|\frac{1}{n_s}\mathbf{Y}^T\mathbf{Y}-\boldsymbol{\Sigma}-\frac{1}{n_t}\mathbf{B}\mathbf{X}^T\mathbf{X}\mathbf{B}^T\right\|_F^2+\frac{\lambda}{2}\|\mathbf{D}\circ\mathbf{B}\|_F^2, \quad [4]$$

where λ is the regularization parameter, $\mathbf{D}=(D_{ij})\in\mathbb{R}^{p\times q}$ is a scaling or weight matrix, and $\circ$ denotes the Hadamard product. By default, λ, which controls the confidence in prior knowledge, is set to 0.0001. For integrating scRNA-seq and scATAC-seq data, D_{ij} is calculated using an exponential function $D_{ij}=\exp\left(d_{ij}/d_0\right)$, where d_{ij} is the distance between the peak j and the transcription start site of gene i, and d_0 is a hyperparameter with a default value of 100 kb (30). This reflects the biological principle that peak-gene regulatory potential decays with increasing genomic distance. For peak-gene pairs separated by more than 200 kb or located on different chromosomes, B_{ij} is constrained to zero. This design primarily aims to improve computational efficiency by restricting the estimation to chromatin regions proximal to each gene, thereby greatly reducing the parameter space. Otherwise, estimating all entries of $\mathbf{B}$ would significantly increase computational complexity. From a biological perspective, this is also reasonable, since gene expression is predominantly influenced by nearby regions. Other forms of prior knowledge, such as eQTL information, can also be incorporated through this penalty term, as detailed in *SI Appendix*, Text.

Estimation. We design an efficient algorithm within the ADMM framework to compute the solution to [**4**]. Let

$$f_i(\boldsymbol{\beta}_i,\Sigma_{ii})=\int\left\{\mathbb{F}_{n_s}^{y_i}(y_i)-\left(\mathbb{F}_{n_t}^{\boldsymbol{\beta}_i^T\mathbf{x}}\star\mathbb{F}^{\epsilon_i}\right)(y_i)\right\}^2 d\mathbb{F}_{n_s}^{y_i}(y_i),$$

and

$$g(\mathbf{B},\boldsymbol{\Sigma})=\frac{w}{2}\left\|\frac{1}{n_s}\mathbf{Y}^T\mathbf{Y}-\boldsymbol{\Sigma}-\frac{1}{n_t}\mathbf{B}\mathbf{X}^T\mathbf{X}\mathbf{B}^T\right\|_F^2+\frac{\lambda}{2}\|\mathbf{D}\circ\mathbf{B}\|_F^2.$$

We assume that $\boldsymbol{\Sigma}$ is diagonal, implying that, given $\mathbf{x}$, the dimensions of $\mathbf{y}$ are independent of each other. Solving the problem [**4**] is equivalent to solving the linearly constrained problem:

$$\min_{\mathbf{B},\mathbf{A},\boldsymbol{\Sigma},\boldsymbol{\Gamma}\succ 0}\sum_{i=1}^{p}f_i(\boldsymbol{\alpha}_i,\gamma_i)+g(\mathbf{B},\boldsymbol{\Sigma})$$
$$\text{s.t.}\begin{pmatrix}\mathbf{A}^T\\ \operatorname{diag}(\boldsymbol{\Gamma})^T\end{pmatrix}-\begin{pmatrix}\mathbf{B}^T\\ \operatorname{diag}(\boldsymbol{\Sigma})^T\end{pmatrix}=0,$$

where $\mathbf{A}\in\mathbb{R}^{p\times q}$, $\boldsymbol{\alpha}_i^T$ is the i-th row of $\mathbf{A}$, and $\boldsymbol{\Gamma}=\operatorname{diag}(\gamma_1,\gamma_2,\ldots,\gamma_p)$ with $\gamma_i>0$ for all $i=1,\ldots,p$. To solve this problem, we consider the augmented Lagrangian in its scaled form:

$$L(\mathbf{B},\boldsymbol{\Sigma},\mathbf{A},\boldsymbol{\Gamma},\boldsymbol{U})=\sum_{i=1}^{p}f_i(\boldsymbol{\alpha}_i,\gamma_i)+g(\mathbf{B},\boldsymbol{\Sigma})+\frac{\rho}{2}\left\|\begin{pmatrix}\mathbf{A}^T\\ \operatorname{diag}(\boldsymbol{\Gamma})^T\end{pmatrix}-\begin{pmatrix}\mathbf{B}^T\\ \operatorname{diag}(\boldsymbol{\Sigma})^T\end{pmatrix}+\boldsymbol{U}\right\|_F^2-\frac{\rho}{2}\|\boldsymbol{U}\|_F^2, \quad [5]$$

where $\boldsymbol{U}\in\mathbb{R}^{(q+1)\times p}$ is the matrix of scaled dual variables, and $\rho>0$ is the penalty parameter with default value of 1. The ADMM algorithm is based on minimizing [**5**] successively over $\{\mathbf{B},\boldsymbol{\Sigma}\}$, $\{\mathbf{A},\boldsymbol{\Gamma}\}$, and $\boldsymbol{U}$. Doing so yields the updates

$$\{\mathbf{B},\boldsymbol{\Sigma}\}^{k+1}=\underset{\mathbf{B},\boldsymbol{\Sigma}}{\operatorname{argmin}}L\left(\mathbf{B},\boldsymbol{\Sigma};\mathbf{A}^k,\boldsymbol{\Gamma}^k,\boldsymbol{U}^k\right), \quad [6]$$

$$\{\mathbf{A},\boldsymbol{\Gamma}\}^{k+1}=\underset{\mathbf{A},\boldsymbol{\Gamma}\succ 0}{\operatorname{argmin}}L\left(\mathbf{A},\boldsymbol{\Gamma};\mathbf{B}^{k+1},\boldsymbol{\Sigma}^{k+1},\boldsymbol{U}^k\right), \quad [7]$$

$$\boldsymbol{U}^{k+1}=\underset{\boldsymbol{U}}{\operatorname{argmax}}L\left(\boldsymbol{U};\mathbf{B}^{k+1},\boldsymbol{\Sigma}^{k+1},\mathbf{A}^{k+1},\boldsymbol{\Gamma}^{k+1}\right). \quad [8]$$

The detailed optimization procedure for each subproblem is described in *SI Appendix*, Text. We alternate through steps [**6**–**8**] repeatedly until convergence. During the optimization process, the objective function steadily converges over iterations, and the model's predictive performance progressively improves with each iteration (*SI Appendix*, Fig. S11). Furthermore, SuperMap demonstrates notable robustness against variations in the hyperparameters λ, ρ, and w (*SI Appendix*, Fig. S12).

Denoising Single-Cell Data. In the SuperMap pipeline, we first fit the model to learn cross-modal mappings. After obtaining $\hat{\mathbf{B}}$, we proceed to impute the missing (unmeasured) modality profiles in unpaired data ($\hat{\mathbf{Y}}^*=\mathbf{X}\hat{\mathbf{B}}^T$) for subsequent analyses. However, single-cell data are highly sparse and noisy, with an inherently low signal-to-noise ratio, which hinders both model fitting and imputation, making it difficult for the model to capture reliable predictive signals. Furthermore, even when $\hat{\mathbf{B}}$ is accurately estimated, direct imputation via $\hat{\mathbf{Y}}^*=\mathbf{X}\hat{\mathbf{B}}^T$ is unreliable due to noise inherent in $\mathbf{X}$. To mitigate these issues, we adopt two widely used denoising strategies for single-cell data–meta-cell partitioning and K-nearest neighbor (KNN) smoothing–each applied to different scenarios.

Meta-cell partitioning is a widely used approach to enhance the signal-to-noise ratio in single-cell data by aggregating closely similar cells into *meta-cells* (42, 61, 62). Specifically, we employed the Leiden algorithm to cluster cells, assuming that cells within each cluster are intrinsically homogeneous and their observed variability is largely due to technical noise rather than biologically meaningful heterogeneity. Subsequently, we aggregated and averaged the cells within each cluster to form meta-cells. This approach effectively enhances the signal-to-noise ratio, but inevitably sacrifices single-cell resolution and lost intra-meta-cell heterogeneity.

In addition to meta-cell partitioning, another widely used strategy for denoising single-cell data is KNN smoothing, which has been successfully applied in various contexts, such as RNA velocity estimation (63, 64). The key idea is to borrow information from neighboring cells to improve the signal-to-noise ratio. Specifically, we first construct the KNN graph in the low-dimensional representation space (e.g., PCA for RNA data) using the *FindNeighbors* function in Seurat pipeline (default $K=50$). From this graph, we obtain each cell's K nearest neighbors and their similarity scores, denoted as s_{ij}, which are calculated as the Jaccard index of neighborhood overlap (details available in the *FindNeighbors* documentation). The s_{ij} values range from 0 to 1, with higher values indicating greater similarity. We then smooth the modality data $\mathbf{X}$ by computing a weighted average of each cell and its neighbors:

$$\mathbf{x}_i=\alpha\mathbf{x}_i+(1-\alpha)\sum_{j\in\mathcal{N}(i)}s_{ij}\mathbf{x}_j,$$

where $\mathcal{N}(i)$ denotes the set of K nearest neighbors for cell i, and $\alpha \in [0, 1]$ controls the smoothing strength (default $\alpha = 0.5$). This approach effectively reduces noise while preserving cell-to-cell heterogeneity.

It is worth noting that meta-cell partitioning and KNN smoothing share a similar concept–both leverage local neighborhood information to improve data quality. Meta-cell partitioning can be viewed as an extreme form of KNN smoothing, where all local neighbors share identical smoothed profiles, which greatly reduces noise but eliminates intraneighborhood heterogeneity. Thus, these two methods represent different trade-offs between denoising strength and resolution. KNN smoothing is more suitable for scenarios that require preserving single-cell heterogeneity, such as data integration, whereas meta-cell partitioning is preferable for tasks that emphasize stable feature relationships, such as regulatory network inference or gene coexpression analysis.

Model Fitting and Missing-Modality Imputation. For estimating cross-modal mappings, since this step focuses on relationships between feature spaces rather than individual cells, we consider both meta-cell data and KNN-smoothed single-cell data to be suitable, as both contain information about the underlying feature mappings. The analysis of the meta-cell graining level further supports this reasoning (*Results*). However, using meta-cell data at this step offers a clear advantage in computational efficiency: the reduced sample size substantially decreases the computational burden, enabling scalability to large datasets. Therefore, we use meta-cell data for model fitting in this step.

For imputation, SuperMap provides options at both meta-cell and single-cell resolution to accommodate different downstream analysis tasks. Specifically, for meta-cell imputation, we perform it at the meta-cell level by applying $\hat{\mathbf{Y}}^*_{meta} = \mathbf{X}_{meta}\hat{\mathbf{B}}^T$. For single-cell imputation, we first smooth $\mathbf{X}$ using the KNN-smoothing procedure described above to obtain $\mathbf{X}_{smooth}$, and then perform imputation as $\hat{\mathbf{Y}}^* = \mathbf{X}_{smooth}\hat{\mathbf{B}}^T$. Because smoothing reduces noise in the predictors, this yields more reliable imputed values while maintaining single-cell heterogeneity. For analyses that require preserving single-cell resolution–such as diagonal integration–we recommend using single-cell imputation. For tasks that primarily focus on feature-level relationships rather than individual cells, such as regulatory network inference or coexpression analysis, meta-cell imputation is preferred, as it provides more stable signals and aligns with strategies commonly adopted in previous studies (15, 42, 65). Moreover, the meta-cell approach is particularly well suited for large-scale cell atlas data, offering substantial advantages in memory efficiency and computational speed (66, 67). Users can choose the appropriate imputation strategy according to their specific research goals and analytical needs.

Diagonal Integration and Regulatory Analysis. We designed a diagonal integration scheme that harnesses the complementary strengths of both SuperMap and Seurat. Specifically, Seurat calculates the gene activity score from ATAC data–based on genomic coordinates–to serve as the linkage for aligning RNA and ATAC modalities. The cross-modal mappings learned by SuperMap enable us to impute the missing RNA profiles for ATAC cells at single-cell level, as described above. Within the Seurat pipeline, we replace the relatively coarse gene activity score with more informative and accurate SuperMap-imputed profiles as linkages, thereby achieving improved modality alignment.

For regulatory analysis, SuperMap first perform imputation at the meta-cell level to generate paired data $\{\hat{\mathbf{Y}}^*, \mathbf{X}\}$. It then calculates the standardized regression coefficients between each gene and its potential regulatory elements (e.g., ATAC peaks), which serve as *regulatory scores* quantifying their cis-regulatory potential. Peak-gene pairs with higher regulatory scores indicate a greater likelihood of exhibiting cis-regulatory interactions.

Trajectory Inference and the Identification of Epigenomic Priming. We applied Monocle3 to infer the cell differentiation trajectory using the integrated latent embeddings of unpaired data. To identify differentiation-dependent genes, we used the *graph_test* function in Monocle3 on RNA cells within the trajectory, considering genes with $q < 0.01$ as statistically significant. To further identify genes with dramatic dynamic changes, we selected the top 500 genes with the strongest effects of trajectory on expression. These genes were subsequently subjected to hierarchical clustering, and their expression patterns were visualized using a heatmap to provide a comprehensive view of their dynamic behavior. To investigate epigenomic priming, we focused on gene clusters demonstrating parabolic expression patterns. We applied natural cubic splines to smooth the ATAC-predicted gene expression and RNA expression data along the inferred trajectory. To quantify the temporal displacement between these two expression profiles, we identified the extrema (points where the first derivative equals zero) of each smoothed curve. The pseudotime difference between the extrema of the ATAC-predicted expression and RNA expression was defined as the temporal displacement. A positive displacement indicates that chromatin state dynamics precede RNA expression, suggesting the presence of epigenomic priming.

Data, Materials, and Software Availability. The 10X Multiome PBMC data can be accessed on 10X Genomics website (https://www.10xgenomics.com/datasets) (68). The 10X BMMC Multiome data is available at GSE194122 (69). The SHARE-seq mouse skin data can be found at GSE140203 (70). The scNMT-seq mouse embryos data can be found at GSE121708 (71). The ASAP-seq PBMC dataset is available at GSE156478 (72). The sciRNA-seq3 dataset of human fetal cerebrum was collected from GSE156793 (73). The sciATAC-seq3 dataset of human fetal cerebrum was collected from GSE149683 (74). The eQTL data can be downloaded from the GTEx v8 website (https://www.gtexportal.org/home/datasets) (75). Details on the benchmarking datasets, data preprocessing steps, evaluation metrics, and baseline methods are provided in *SI Appendix*, Text. The SuperMap software is available at https://github.com/chaodeng-aca/SuperMap (76).

ACKNOWLEDGMENTS. This research was supported in part by the National Natural Science Foundation of China (12571306, 12331009, 12222111, and 12201219), Shanghai Key Program of Computational Biology (23JS1400500 and 23JS1400800), the Fundamental Research Funds for the Central Universities, and the Neil Shen's Shanghai Jiao Tong University Medical Research Fund of Shanghai Jiao Tong University.

Author affiliations: [a]Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, China; [b]Shanghai Jiao Tong University-Yale Joint Center for Biostatistics and Data Science, National Center for Translational Medicine, Shanghai Jiao Tong University, Shanghai 200240, China; [c]Department of Statistics and Key Laboratory of Scientific and Engineering Computing (Ministry of Education) and Shanghai Frontier Science Center of Modern Analysis, School of Mathematical Sciences, Shanghai Jiao Tong University, Shanghai 200240, China; [d]Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT 06511; and [e]Key Laboratory of Advanced Theory and Application in Statistics and Data Science (Ministry of Education), School of Statistics and Academy of Statistics and Interdisciplinary Sciences, East China Normal University, Shanghai 200062, China

1. D. A. Cusanovich *et al.*, Multiplex single-cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
2. C. A. Lareau *et al.*, Droplet-based combinatorial indexing for massive-scale single-cell chromatin accessibility. *Nat. Biotechnol.* **37**, 916–924 (2019).
3. C. Luo *et al.*, Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
4. S. Picelli *et al.*, Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
5. E. Z. Macosko *et al.*, Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
6. B. Budnik, E. Levy, G. Harmange, N. Slavov, SCoPE-MS: Mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome Biol.* **19**, 1–12 (2018).
7. S. Chen, B. B. Lake, K. Zhang, High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
8. S. Ma *et al.*, Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
9. S. J. Clark *et al.*, scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
10. M. Stoeckius *et al.*, Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
11. C. Zhu, S. Preissl, B. Ren, Single-cell multimodal omics: The power of many. *Nat. Methods* **17**, 11–14 (2020).
12. T. Stuart, R. Satija, Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
13. Z. Miao, B. D. Humphreys, A. P. McMahon, J. Kim, Multi-omics integration in the age of million single-cell data. *Nat. Rev. Nephrol.* **17**, 710–724 (2021).

14. Y. Jiang *et al*., Nonparametric single-cell multiomic characterization of trio relationships between transcription factors, target genes, and cis-regulatory regions. *Cell Syst.* **13**, 737–751 (2022).
15. Q. Yuan, Z. Duren, Inferring gene regulatory networks from single-cell multiome data using atlas-scale external data. *Nat. Biotechnol.* **43**, 1–11 (2024).
16. R. Argelaguet, A. S. Cuomo, O. Stegle, J. C. Marioni, Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* **39**, 1202–1215 (2021).
17. Y. Xu, R. P. McCord, Diagonal integration of multimodal single-cell data: Potential pitfalls and paths forward. *Nat. Commun.* **13**, 3505 (2022).
18. T. Stuart *et al*., Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
19. J. Dou *et al*., Bi-order multimodal integration of single-cell data. *Genome Biol.* **23**, 112 (2022).
20. Y. Lin *et al*., scJoint integrates atlas-scale single-cell RNA-seq and ATAC-seq data with transfer learning. *Nat. Biotechnol.* **40**, 703–710 (2022).
21. Z. J. Cao, G. Gao, Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. *Nat. Biotechnol.* **40**, 1458–1466 (2022).
22. Z. Duren *et al*., Integrative analysis of single-cell genomics data by coupled nonnegative matrix factorizations. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 7723–7728 (2018).
23. Y. Hao *et al*., Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
24. S. Ghazanfar, C. Guibentif, J. C. Marioni, Stabilized mosaic single-cell data integration using unshared features. *Nat. Biotechnol.* **42**, 284–292 (2024).
25. K. E. Wu, K. E. Yost, H. Y. Chang, J. Zou, BABEL enables cross-modality translation between multiomic profiles at single-cell resolution. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2023070118 (2021).
26. Y. Cao *et al*., scButterfly: A versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. *Nat. Commun.* **15**, 2973 (2024).
27. E. H. Davidson *et al*., A genomic regulatory network for development. *Science* **295**, 1669–1678 (2002).
28. P. Badia-i Mompel *et al*., Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
29. S. Boyd *et al*., Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.* **3**, 1–122 (2011).
30. Q. Yuan, Z. Duren, Integration of single-cell multi-omics data by regression analysis on unpaired observations. *Genome Biol.* **23**, 160 (2022).
31. M. Y. Lee, K. H. Kaestner, M. Li, Benchmarking algorithms for joint integration of unpaired and paired single-cell RNA-seq and ATAC-seq data. *Genome Biol.* **24**, 244 (2023).
32. M. Luecken *et al*., "A sandbox for prediction and integration of DNA, RNA, and proteins in single cells" in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, J. Vanschoren, S. Yeung., Eds. (Curran Associates, 2021), vol. 1.
33. R. Argelaguet *et al*., Multi-omics profiling of mouse gastrulation at single-cell resolution. *Nature* **576**, 487–491 (2019).
34. E. P. Mimitou *et al*., Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. *Nat. Biotechnol.* **39**, 1246–1258 (2021).
35. T. Stuart, A. Srivastava, S. Madad, C. A. Lareau, R. Satija, Single-cell chromatin state analysis with Signac. *Nat. Methods* **18**, 1333–1341 (2021).
36. C. Wang *et al*., Integrative analyses of single-cell transcriptome and regulome using MAESTRO. *Genome Biol.* **21**, 1–28 (2020).
37. L. D. Moore, T. Le, G. Fan, DNA methylation and its basic function. *Neuropsychopharmacology* **38**, 23–38 (2013).
38. A. L. Mattei, N. Bailly, A. Meissner, DNA methylation: A historical perspective. *Trends Genet.* **38**, 676–707 (2022).
39. Y. Uzun, H. Wu, K. Tan, Predictive modeling of single-cell DNA methylome data enhances integration with transcriptome data. *Genome Res.* **31**, 101–109 (2021).
40. J. Liu, Y. Huang, R. Singh, J. P. Vert, W. S. Noble, "Jointly embedding multiple single-cell omics measurements" in *19th International Workshop on Algorithms in Bioinformatics (WABI 2019)*, K. T. Huber, D. Gusfield, Eds. (Dagstuhl Publishing, 2019), vol. 143, pp. 10:1–10:13.
41. M. D. Luecken *et al*., Benchmarking atlas-level data integration in single-cell genomics. *Nat. Methods* **19**, 41–50 (2022).
42. S. Persad *et al*., SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757 (2023).
43. GTEx Consortium, Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
44. R. S. Bystry, V. Aluvihare, K. A. Welch, M. Kallikourdis, A. G. Betz, B cells and professional APCs recruit regulatory T cells via CCL4. *Nat. Immunol.* **2**, 1126–1132 (2001).
45. W. D. Raymond, G. Ø. Eilertsen, S. Shanmugakumar, J. C. Nossent, The impact of cytokines on the health-related quality of life in patients with systemic lupus erythematosus. *J. Clin. Med.* **8**, 857 (2019).
46. S. Ghafouri-Fard, M. Shahir, M. Taheri, A. Salimi, A review on the role of chemokines in the pathogenesis of systemic lupus erythematosus. *Cytokine* **146**, 155640 (2021).
47. J. M. Granja *et al*., ArchR is a scalable software package for integrative single-cell chromatin accessibility analysis. *Nat. Genet.* **53**, 403–411 (2021).
48. M. T. Weirauch *et al*., Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* **158**, 1431–1443 (2014).
49. J. Cao *et al*., A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
50. S. Domcke *et al*., A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
51. A. A. Pollen *et al*., Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55–67 (2015).
52. E. R. Thomsen *et al*., Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods* **13**, 87–93 (2016).
53. J. Cao *et al*., The single-cell transcriptional landscape of mammalian organogenesis. *Nature* **566**, 496–502 (2019).
54. C. Ernst, M. Jefri, Epigenetic priming in neurodevelopmental disorders. *Trends Mol. Med.* **27**, 1106–1114 (2021).
55. M. Meijer *et al*., Epigenomic priming of immune genes implicates oligodendroglia in multiple sclerosis susceptibility. *Neuron* **110**, 1193–1210 (2022).
56. J. Olusakin *et al*., Implication of 5-HT7 receptor in prefrontal circuit assembly and detrimental emotional effects of SSRIs during development. *Neuropsychopharmacology* **45**, 2267–2277 (2020).
57. X. Ruan *et al*., Progenitor cell diversity in the developing mouse neocortex. *Proc. Natl. Acad. Sci. U.S.A.* **118**, e2018866118 (2021).
58. R. N. Delgado *et al*., Individual human cortical progenitors can produce excitatory and inhibitory neurons. *Nature* **601**, 397–403 (2022).
59. F. Balabdaoui, C. R. Doss, C. Durot, Unlinked monotone regression. *J. Mach. Learn. Res.* **22**, 1–60 (2021).
60. M. Azadkia, F. Balabdaoui, Linear regression with unmatched data: A deconvolution perspective. *J. Mach. Learn. Res.* **25**, 1–55 (2024).
61. P. Liu, J. J. Li, mcRigor: A statistical method to enhance the rigor of metacell partitioning in single-cell data analysis. *Nat. Commun.* **16**, 8602 (2025).
62. M. Bilous, L. Hérault, A. A. Gabriel, M. Teleman, D. Gfeller, Building and analyzing metacells in single-cell genomics data. *Mol. Syst. Biol.* **20**, 1–23 (2024).
63. V. Bergen, M. Lange, S. Peidli, F. A. Wolf, F. J. Theis, Generalizing RNA velocity to transient cell states through dynamical modeling. *Nat. Biotechnol.* **38**, 1408–1414 (2020).
64. A. Gayoso *et al*., Deep generative modeling of transcriptional dynamics for RNA velocity analysis in single cells. *Nat. Methods* **21**, 50–59 (2024).
65. Y. Baran *et al*., MetaCell: Analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 1–19 (2019).
66. L. Zheng *et al*., Pan-cancer single-cell landscape of tumor-infiltrating T cells. *Science* **374**, abe6474 (2021).
67. O. Ben-Kiki *et al*., MCProj: Metacell projection for interpretable and quantitative use of transcriptional atlases. *Genome Biol.* **24**, 220 (2023).
68. 10x Genomics, Datasets. 10x Genomics. https://www.10xgenomics.com/datasets. Deposited 3 May 2021.
69. M. Luecken *et al*., A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE194122. Deposited 24 January 2022.
70. S. Ma *et al*., Integrative single-cell chromatin and transcriptome profiling uncovers cell-type specific regulatory interactions. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE140203. Deposited 23 October 2020.
71. R. Argelaguet *et al*., Multi-omics profiling of mouse gastrulation at single cell resolution. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121708. Deposited 30 September 2019.
72. E. P. Mimitou *et al*., High throughput joint profiling of chromatin accessibility and protein levels in single cells. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156478. Deposited 8 September 2020.
73. J. Cao *et al*., A human cell atlas of fetal gene expression. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE156793. Deposited 1 November 2020.
74. S. Domcke *et al*., A human cell atlas of fetal chromatin accessibility. GEO. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE149683. Deposited 1 November 2020.
75. The GTEx Consortium, Download Open Access Datasets. GTEx Portal. https://www.gtexportal.org/home/datasets. Deposited 9 February 2023.
76. C. Deng *et al*., SuperMap. GitHub. https://github.com/chaodeng-aca/SuperMap. Deposited 14 November 2025.