Multi-omics integration and regulatory inference for unpaired single-cell data with a graph-linked unified embedding framework

Zhi-Jie Cao[1], Ge Gao[1,*]

[1] Biomedical Pioneering Innovation Center (BIOPIC), Beijing Advanced Innovation Center for Genomics (ICG), Center for Bioinformatics (CBI), and State Key Laboratory of Protein and Plant Gene Research at School of Life Sciences, Peking University, Beijing, 100871, China

* To whom correspondence should be addressed. Tel: +86-010-62755206; Email: gaog@mail.cbi.pku.edu.cn

Abstract

With the ever-increasing amount of single-cell multi-omics data accumulated during the past years, effective and efficient computational integration is becoming a serious challenge. One major obstacle of unpaired multi-omics integration is the feature discrepancies among omics layers. Here, we propose a computational framework called GLUE (graph-linked unified embedding), which utilizes accessible prior knowledge about regulatory interactions to bridge the gaps between feature spaces. Systematic benchmarks demonstrated that GLUE is accurate, robust and scalable. We further employed GLUE for various challenging tasks, including triple-omics integration, model-based regulatory inference and multi-omics human cell atlas construction (over millions of cells) and found that GLUE achieved superior performance for each task. As a generalizable framework, GLUE features a modular design that can be flexibly extended and enhanced for new analysis tasks. The full package is available online at https://github.com/gao-lab/GLUE for the community.

Introduction

Recent technological advances in single-cell sequencing have enabled the probing of regulatory maps through multiple omics layers, such as chromatin accessibility (scATAC-seq[1, 2]), DNA methylation (snmC-seq[3], sci-MET[4]) and the transcriptome (scRNA-seq[5, 6]), offering a unique opportunity to unveil the underlying regulatory bases for the functionalities of diverse cell types[7]. While simultaneous assays are emerging recently[8-11], different omics are usually measured independently and produce unpaired data, which calls for effective and efficient *in silico* multi-omics integration[12, 13].

Computationally, one major obstacle faced when integrating unpaired multi-omics data is the distinct feature spaces of different modalities (e.g., accessible chromatin regions in scATAC-seq vs. genes in scRNA-seq)[14]. A quick fix is to convert multimodality data into one common feature space based on prior information and apply single-omics data integration methods[15, 16]. Such explicit "feature conversion" is straightforward, but has been reported to result in significant information loss[17]. Algorithms based on coupled matrix factorization circumvent explicit conversion but hardly handle more than two omics layers[18, 19]. An alternative option is to match cells from different omics layers via nonlinear manifold alignment, which removes the requirement of prior knowledge completely and could reduce inter-modality information loss in theory[20, 21]; however, this technique has mostly been applied to continuous, trajectory-like manifolds rather than atlases.

The ever-increasing volume of data is another serious challenge[22]. Recently developed technologies can routinely generate datasets at the scale of millions of cells[23-25], whereas current integration methods have only been applied to datasets with much smaller volumes[15, 16, 18-21]. To catch up with the growth in data throughput, computational integration methods should be designed with scalability in mind.

Hereby, we introduce GLUE (graph-linked unified embedding), a modular framework for integrating unpaired single-cell multi-omics data and inferring regulatory interactions simultaneously. By modeling the regulatory interactions across omics layers explicitly, GLUE bridges the gaps between various omics-specific feature spaces in a biologically intuitive manner. Systematic benchmarks and case studies demonstrate that GLUE is accurate, robust and scalable for heterogeneous single-cell multi-omics data. Furthermore, GLUE is designed as a generalizable framework that allows for easy

extension and quick adoption to particular scenarios in a modular manner. GLUE is publicly accessible at https://github.com/gao-lab/GLUE.

Results

Integrating unpaired single-cell multi-omics data via graph-guided embeddings

Inspired by previous works, we model cell states as low-dimensional cell embeddings learned through variational autoencoders[26, 27]. Given their intrinsic differences in biological nature and assay technology, each omics layer is equipped with a separate autoencoder that uses a probabilistic generative model tailored to the layer-specific feature space (Fig. 1, Methods).

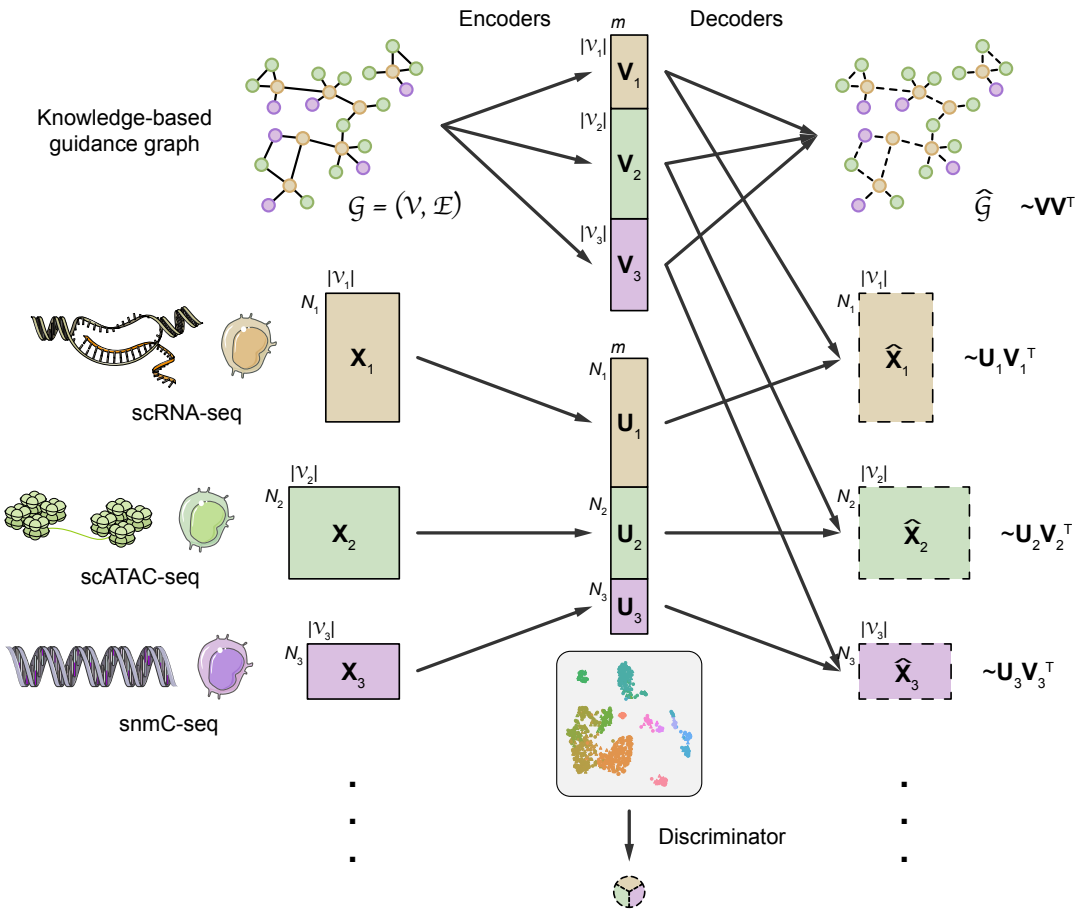


Fig. 1 Architecture of the GLUE framework.
GLUE employs omics-specific variational autoencoders to learn low-dimensional cell embeddings from each omics layer. The data dimensionality and generative distribution can differ across omics layers, but the cell embedding dimensions are shared. A graph variational autoencoder is used to learn feature embeddings from the prior knowledge-based guidance graph; these embeddings are then used as data decoder parameters. The feature embeddings effectively link the omics-specific autoencoders to ensure a consistent embedding orientation. Last, an omics discriminator is employed to align the cell embeddings of different omics layers via adversarial learning.

Taking advantage of prior biological knowledge, we propose the use of a knowledge-based graph ("guidance graph") that explicitly models cross-layer regulatory interactions for linking layer-specific feature spaces; the vertices in the graph correspond to the features of different omics layers, and edges represent signed regulatory interactions. For example, when integrating scRNA-seq and scATAC-seq data, the vertices are genes and accessible chromatin regions (i.e., ATAC peaks), and a positive edge can be connected between an accessible region and its putative downstream gene. Then, adversarial multimodal alignment is performed as an iterative optimization procedure, guided by feature embeddings encoded from the graph[28] (Fig. 1, Methods). Notably, when the iterative process converges, the graph can be refined with inputs from the alignment procedure and used for data-oriented regulatory inference (see below for more details).

Systematic benchmarks demonstrate superior alignment accuracy and robustness over existing methods

We first benchmarked GLUE against multiple popular unpaired multi-omics integration methods[15, 16, 20, 29] using gold-standard datasets generated by recent simultaneous scRNA-seq and scATAC-seq technologies[8, 9, 30].

At the cell type level, an integration method should match the corresponding cell types from different omics layers, producing cell embeddings where the cell types are clearly distinguishable and the omics layers are well mixed. Compared to other methods, GLUE achieved overall the highest cell type resolution (as quantified by mean average precision) and layer mixing (as quantified by the Seurat alignment score[31]) simultaneously (Fig. 2a); these results were also validated by UMAP visualization of the aligned cell embeddings (Supplementary Fig. 1-3).

An optimal integration method should produce accurate alignments not only at the cell type level but also at finer scales. Exploiting the ground truth cell-to-cell correspondence between scRNA-seq and scATAC-seq, we further quantified single-cell level alignment error via the FOSCTTM (fraction of samples closer than the true match) metric[32]. On all three datasets, GLUE achieved the lowest FOSCTTM, decreasing the alignment error by large margins compared to the second-best method on each dataset (Fig. 2b, the decreases were 3.6-fold for SNARE-seq, 1.74-fold for SHARE-seq, and 1.4-fold for 10x Multiome).

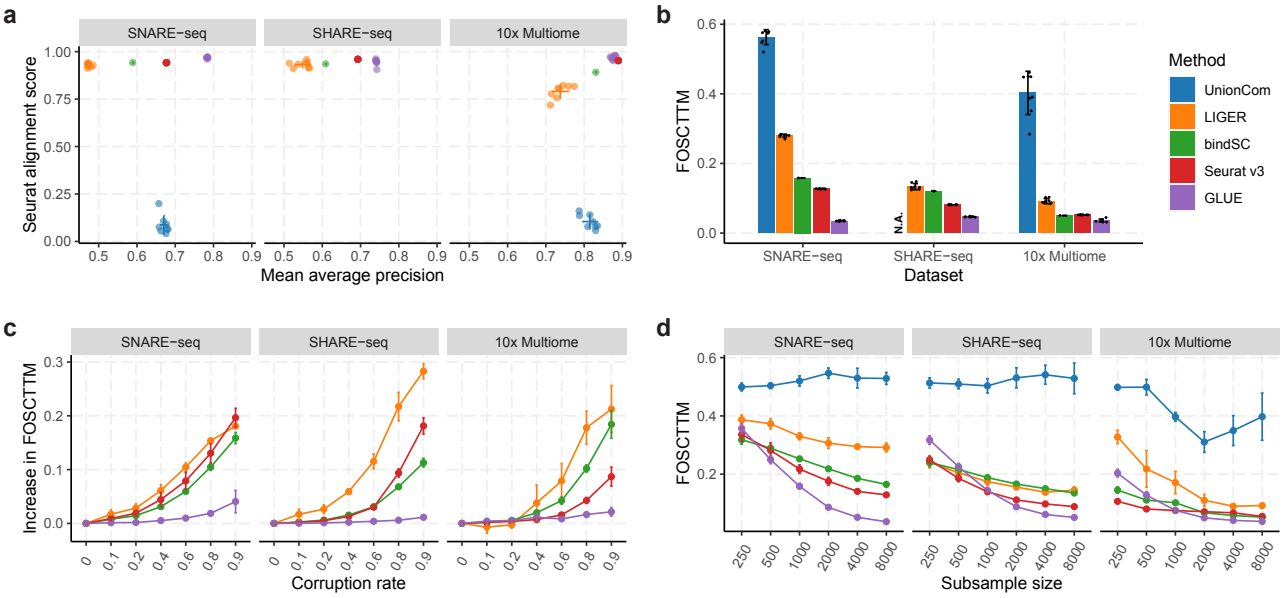


Fig. 2 Systematic benchmarks on gold-standard datasets.
a, Cell type resolution (quantified by mean average precision) vs. omics layer mixing (quantified by Seurat alignment score) for different integration methods. **b**, Single-cell level alignment error (quantified by FOSCTTM) of different integration methods. UnionCom failed to run on the SHARE-seq dataset due to memory overflow. **c**, Increases in FOSCTTM at different prior knowledge corruption rates for integration methods that rely on prior feature interactions. **d**, FOSCTTM values of different integration methods on subsampled datasets. The error bars indicate mean ± s.d.

During the evaluation described above, we adopted a standard schema (ATAC peaks were linked to RNA genes if they overlapped in the gene body or proximal promoter regions) to construct the guidance graph for GLUE and to perform feature conversion for other conversion-based methods. Given that our current knowledge about the regulatory interactions is still far from prefect, a useful integration method must be robust to such inaccuracies. Thus, we further assessed the methods' robustness to corruption of regulatory interactions by randomly replacing varying fractions of existing interactions with nonexistent ones. For all three datasets, GLUE exhibited the smallest performance changes even at high corruption rates (Fig. 2c), suggesting its superior robustness.

Given its neural network-based nature, GLUE may suffer from undertraining when working with small datasets. Thus, we repeated the evaluations using subsampled datasets of various sizes. GLUE remained the top-ranking method with as few as 2,000 cells, but the alignment error increased more steeply when the data volume decreased to less than 1,000 cells (Fig. 2d). Additionally, we also noted that the performance of GLUE was robust for a wide range of hyperparameter settings (Supplementary Fig. 4).

GLUE enables effective triple-omics integration

Benefitting from a modular design and scalable adversarial alignment, GLUE readily extends to more than two omics layers. As a case study, we used GLUE to integrate three distinct omics layers of neuronal cells in the adult mouse cortex, including gene expression[33], chromatin accessibility[34], and DNA methylation[3].

Unlike chromatin accessibility, gene body DNA methylation generally shows a negative correlation with gene expression in neuronal cells[35]. GLUE natively supports the mixture of regulatory effects by modeling edge signs in the guidance graph. Such a strategy avoids data inversion, which is required by previous methods[16] and can break data sparsity and the underlying distribution. For the triple-omics guidance graph, we linked gene body mCH and mCG levels to genes via negative edges, while the positive edges between accessible regions and genes remained the same.

The GLUE alignment successfully revealed a shared manifold of cell states across the three omics layers (Fig. 3a-d). We observed highly significant marker overlap (Fig. 3e, three-way Fisher's exact test[36], FDR $< 2\times10^{-15}$) for 12 out of the 14 mapped cell types (Supplementary Fig. 5, 6, Methods), indicating reliable alignment. Interestingly, we found that GLUE alignment helped improve the effects of cell typing in all omics layers, including the further partitioning of the scRNA-seq "MGE" cluster into *Pvalb*$^+$ ("mPv") and *Sst*$^+$ ("mSst") subtypes (highlighted with green circles/flows in Fig. 3, Supplementary Fig. 5), the partitioning of the scRNA-seq "CGE" cluster and scATAC-seq "Vip" cluster into *Vip*$^+$ ("mVip") and *Ndnf*$^+$ ("mNdnf") subtypes (highlighted with dark blue circles/flows in Fig. 3, Supplementary Fig. 5), and the identification of snmC-seq "mDL-3" cells and a subset of scATAC-seq "L6 IT" cells as claustrum cells (highlighted with light blue circles/flows in Fig. 3, Supplementary Fig. 5).

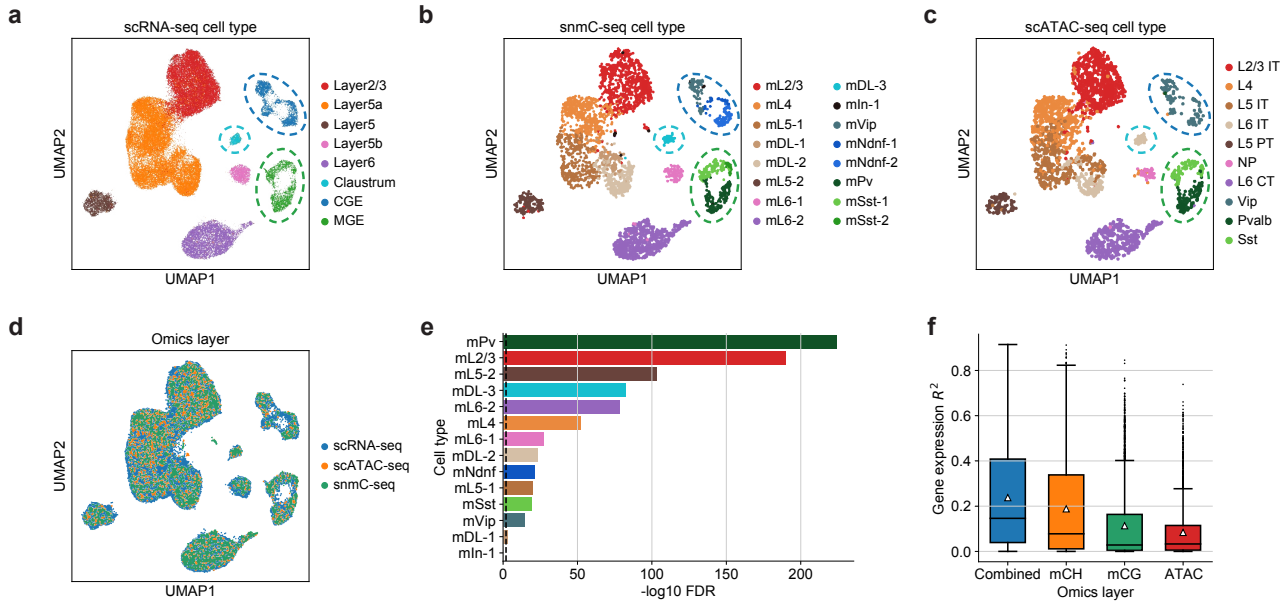


Fig. 3 Triple-omics integration of the mouse cortex.
a-c, UMAP visualizations of the integrated cell embeddings for **a**, scRNA-seq, **b**, snmC-seq, and **c**, scATAC-seq, colored by the original cell types. Cells aligning with "mPv" and "mSst" are highlighted with green circles. Cells aligning with "mNdnf" and "mVip" are highlighted with dark blue circles. Cells aligning with "mDL-3" are highlighted with light blue circles. **d**, UMAP visualizations of the integrated cell embeddings for all cells, colored by omics layers. **e**, Significance of marker gene overlap for each cell type across all three omics layers (three-way Fisher's exact test[36]). The dashed vertical line indicates that FDR = 0.01. We observed highly significant marker overlap (FDR $< 2 \times 10^{-15}$) for 12 out of the 14 cell types, indicating reliable alignment. For the remaining 2 cell types, "mDL-1" had marginally significant marker overlap with FDR = 0.001, while the "mIn-1" cells in snmC-seq did not properly align with the scRNA-seq or scATAC-seq cells. **f**, Coefficient of determination (R^2) for predicting gene expression based on each epigenetic layer as well as the combination of all layers. The box plots indicate the medians (centerlines), means (triangles), 1st and 3rd quartiles (hinges), and minima and maxima (whiskers).

Such triple-omics integration also sheds light on the quantitative contributions of different epigenetic regulation mechanisms (Methods). Among mCH, mCG and chromatin accessibility, we found that the mCH level had the highest predictive power for gene expression in cortical neurons (average R^2 = 0.188). When all epigenetic layers were considered, the expression predictability increased further (average R^2 = 0.238), suggesting the presence of nonredundant contributions (Fig. 3f). Among the neurons of different layers, DNA methylation (especially mCH) exhibited slightly higher predictability for gene expression in deeper layers than in superficial layers, whereas the reverse situation held for chromatin accessibility (Supplementary Fig. 7a). Across all genes, the predictability of gene expression was generally correlated among the different epigenetic layers (Supplementary Fig. 7b). We also observed varying associations with gene characteristics. For example, mCH had higher expression predictability for longer genes, which was consistent with previous studies[16, 37], while chromatin accessibility contributed more to genes with higher expression variability (Supplementary Fig. 7c).

Model-based regulatory inference with GLUE

The incorporation of a graph explicitly modeling regulatory interactions in GLUE further enables a Bayesian-like approach that combines prior knowledge and observed data for posterior regulatory inference. Specifically, since the feature embeddings are designed to reconstruct the knowledge-based guidance graph and single-cell multi-omics data simultaneously (Fig. 1), their cosine similarities should reflect information from both aspects, which we adopt as "regulatory scores".

As a demonstration, we employed the official PBMC (peripheral blood mononuclear cell) Multiome dataset from 10x[30] and fed it to GLUE as unpaired scRNA-seq and scATAC-seq data (Methods). To capture remote cis-regulatory interactions, we employed a long-range guidance graph connecting ATAC peaks and RNA genes within 150 kb windows weighted by a power-law function that models chromatin contact probability[38, 39]. Visualization of cell embeddings confirmed that the GLUE alignment was correct and accurate (Supplementary Fig. 8a, b). As expected, we found that the regulatory score was negatively correlated with genomic distance (Fig. 4a) and positively correlated with the empirical peak-gene correlation (computed with paired cells, Fig. 4b), with robustness across different random seeds (Supplementary Fig. 8c).

To further assess whether the score reflected actual cis-regulatory interactions, we compared it with external evidence, including pcHi-C[40] and eQTL[41]. The GLUE regulatory score was higher for pcHi-C-supported peak-gene pairs in all distance ranges (Fig. 4a) and was a better predictor of pcHi-C interactions than empirical peak-gene correlations (Fig. 4b, c), as well as Cicero[39], the coaccessibility-based regulatory prediction method (Fig. 4c). The same held for eQTL (Supplementary Fig. 8d-f).

The GLUE framework also allows additional regulatory evidence, such as pcHi-C, to be incorporated intuitively via the guidance graph. Thus, we further trained new models with a composite guidance graph containing distance-weighted interactions as well as pcHi-C- and eQTL-supported interactions (Supplementary Fig. 9). While the multi-omics alignment was insensitive to these changes, the GLUE-derived TF-target gene network (Methods) showed more significant agreement with manually curated connections in the TRRUST v2 database[42] than individual evidence-based networks (Supplementary Fig. 9e, Supplementary Fig. 10, Supplementary Table 3).

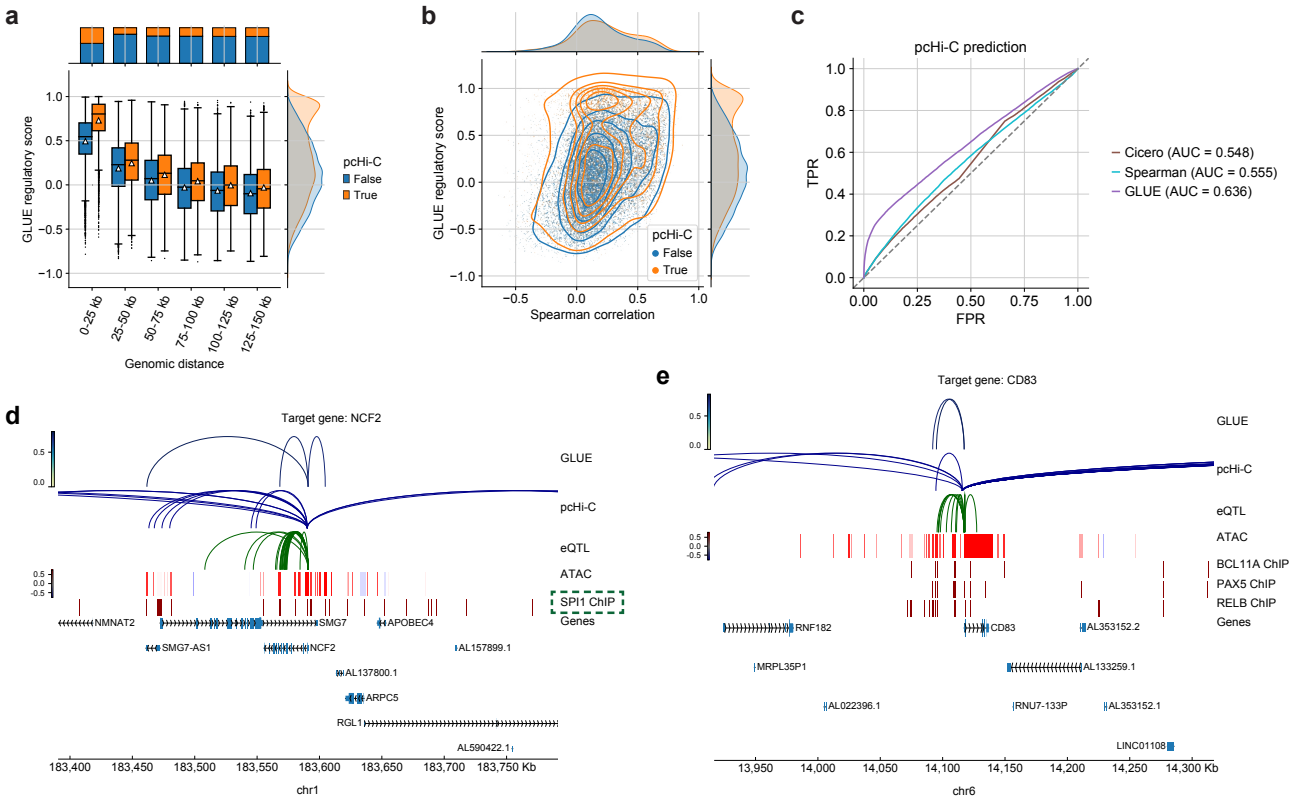


Fig. 4 Model-based regulatory inference in PBMC.
a, GLUE regulatory scores for peak-gene pairs across different genomic ranges, grouped by whether they had pcHi-C support. The box plots indicate the medians (centerlines), means (triangles), 1st and 3rd quartiles (hinges), and minima and maxima (whiskers). **b**, Comparison between the GLUE regulatory scores and the empirical peak-gene correlations computed on paired cells. Peak-gene pairs are colored by whether they had pcHi-C support. **c**, ROC (receiver operating characteristic) curves for predicting pcHi-C interactions based on different peak-gene association scores. **d**, **e**, GLUE-identified cis-regulatory interactions for **d**, *NCF2*, and **e**, *CD83*, along with individual regulatory evidence. SPI1 (highlighted with a green box) is a known regulator of *NCF2*.

Interestingly, we noticed that the GLUE-inferred cis-regulatory interactions could provide new hints about the regulatory mechanisms of known TF-target pairs. For example, SPI1 is a known regulator of the *NCF2* gene, and both are highly expressed in monocytes (Supplementary Fig. 11a, b). GLUE identified three remote regulatory peaks for *NCF2* with various pieces of evidence, i.e., ~120 kb downstream, ~25 kb downstream, and ~20 kb upstream from the TSS (transcription start site) (Fig. 4d), all of which were bound by SPI1. Meanwhile, most putative regulatory interactions were previously unknown. For example, *CD83* was linked with two regulatory peaks (~25 kb upstream from the TSS), which were enriched for the binding of three TFs (BCL11A, PAX5, and RELB; Fig. 4e). While *CD83* was highly expressed in both monocytes and B cells, the inferred TFs showed more constrained expression patterns (Supplementary Fig. 11c-f), suggesting that its active regulators might differ per cell type. Supplementary Fig. 12 shows more examples of GLUE-inferred regulatory interactions.

Atlas-scale integration over millions of cells with GLUE

As technologies continue to evolve, the throughput of single-cell experiments is constantly increasing. Recent studies have generated human cell atlases for gene expression[24] and chromatin accessibility[25] containing millions of cells. The integration of these atlases poses a significant challenge to computational methods due to the sheer volume of data, extensive heterogeneity, low coverage per cell, and unbalanced cell type compositions, and has yet to be accomplished at the single-cell level.

Implemented as parametric neural networks with minibatch optimization, GLUE delivers superior scalability with a sublinear time cost, promising its applicability at the atlas scale, which is barely possible for other methods (Supplementary Fig. 13a). Using an efficient multistage training strategy for GLUE (Methods), we successfully integrated the gene expression and chromatin accessibility data into a unified multi-omics human cell atlas (Fig. 5).

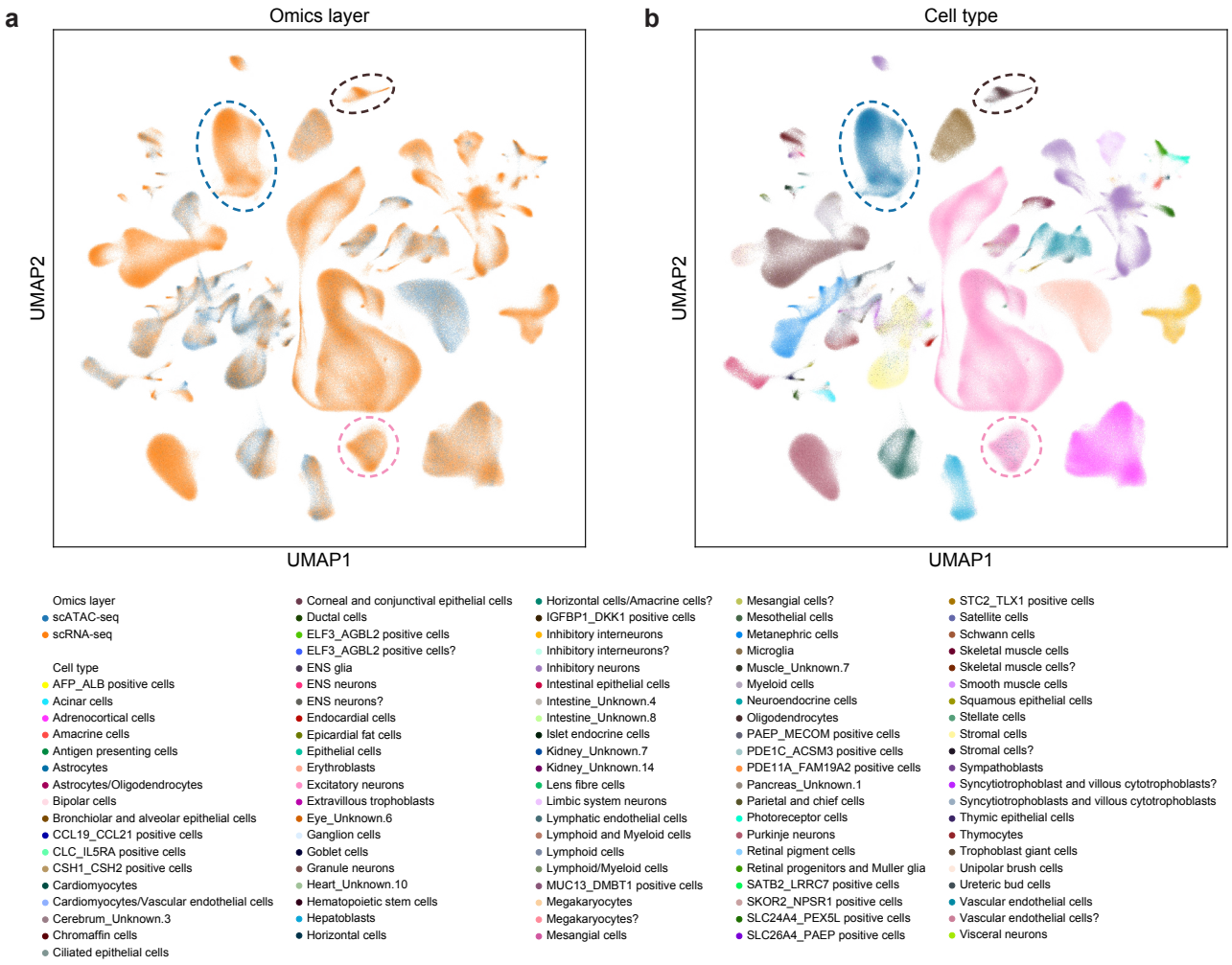


Fig. 5 Integration of a multi-omics human cell atlas.
UMAP visualizations of the integrated cell embeddings, colored by **a**, omics layers, and **b**, cell types. The pink circles highlight cells labeled as "Excitatory neurons" in scRNA-seq but "Astrocytes" in scATAC-seq. The blue circles highlight cells labeled as "Astrocytes" in scRNA-seq but "Astrocytes/Oligodendrocytes" in scATAC-seq. The brown circles highlight cells labeled as "Oligodendrocytes" in scRNA-seq but "Astrocytes/Oligodendrocytes" in scATAC-seq.

While the aligned atlas was largely consistent with the original annotations[25] (Supplementary Fig. 13c-e), we also noticed several discrepancies. For example, cells originally annotated as "Astrocytes" in scATAC-seq were aligned to an "Excitatory neurons" cluster in scRNA-seq (highlighted with pink circles/flows in Supplementary Fig. 13). Further inspection revealed that canonical radial glial (RG) markers such as *PAX6*, *HES1*, and *HOPX*[43, 44] were actively transcribed in this cluster, both in the RNA and ATAC domain (Supplementary Fig. 14), with chromatin priming[9] also detected at both neuronal and glial markers (Supplementary Fig. 15-17), suggesting that the cluster consists of multipotent neural progenitors (likely RGs) rather than excitatory neurons or astrocytes as originally annotated. GLUE-based integration also resolved several scATAC-seq clusters that were ambiguously annotated. For example, the "Astrocytes/Oligodendrocytes" cluster was split into two halves and aligned to the "Astrocytes" and "Oligodendrocytes" clusters of scRNA-seq (highlighted with blue and brown circles/flows in Supplementary Fig. 13, respectively), which was also supported by marker expression and accessibility (Supplementary Fig. 16, 17). These results demonstrate the unique value of atlas-scale multi-omics integration.

Discussion

Combining omics-specific autoencoders with graph-based coupling and adversarial alignment, we designed and implemented the GLUE framework for unpaired single-cell multi-omics data integration with superior accuracy and robustness. By modeling regulatory interactions across omics layers explicitly, GLUE uniquely supports model-based regulatory inference for unpaired multi-omics datasets, exhibiting even higher reliability than regular correlation analysis on paired datasets (notably, in a Bayesian interpretation, the GLUE regulatory inference can be seen as a posterior estimate, which can be continuously refined upon the arrival of new data). Furthermore, benefitting from a neural network-based design, GLUE enables notable scalability for whole-atlas alignment over millions of unpaired cells, which remains a serious challenge for *in silico* integration. In fact, we also attempted to perform integration using the popular Seurat v3 method (Methods), but the result was far from optimal (Supplementary Fig. 18).

Unpaired multi-omics integration, also referred to as diagonal integration[14], shares some conceptual similarities with batch effect correction[45], as both call for the alignment of unpaired cells in certain data representations. Nonetheless, the former is significantly more challenging because of the distinct, omics-specific feature spaces. While completely unsupervised multi-omics integration has been proposed[20, 21], such an approach is exceedingly difficult and has largely been limited to aligning continuous trajectories. For general-case multi-omics integration, additional prior knowledge is necessary. At the omics feature level, presumed feature interactions have been used via feature conversion[15, 16] or coupled matrix factorization[18, 19]. At the cell level, known cell types have also been used via (semi)supervised learning[46, 47], but this approach incurs substantial limitations in terms of applicability since such supervision is typically unavailable and in many cases serves as the purpose of multi-omics integration *per se*[25]. Notably, one of these methods was proposed with a similar autoencoder architecture and adversarial alignment[47], but it relied on matched cell types or clusters to orient the alignment. In fact, GLUE shares more conceptual similarity with the coupled matrix factorization methods, but with superior accuracy, robustness and scalability, which mostly benefits from its deep generative model-based design.

As a generalizable framework, GLUE features a modular design, where the data and graph autoencoders are independently configurable.

- The data autoencoders in GLUE are customizable with appropriate generative models that conform to omics-specific data distributions. In the current work, we used the negative binomial distribution for scRNA-seq and scATAC-seq, and the zero-inflated log-normal distribution for snmC-seq (Methods). Nevertheless, generative distributions can be easily reconfigured to accommodate other omics layers, such as protein abundance[48] and histone modification[49], and to adopt new advances in data modeling techniques[50].
- The guidance graphs used in GLUE have currently been limited to multipartite graphs, containing only edges between the features of different layers. Nonetheless, graphs, as intuitive and flexible representations of regulatory knowledge, can embody more complex regulatory patterns, including within-modality interactions, non-feature vertices, and multi-relations. Beyond canonical graph convolution, more advanced graph neural network architectures[51-53] may also be adopted to extract richer information from the regulatory graph.

Recent advances in experimental multi-omics technologies have increased the availability of paired data[8-11, 30]. While most of the current simultaneous multi-omics protocols still suffer from lower data quality or throughput than that of single-omics methods[54], paired cells can be highly informative in

anchoring different omics layers and should be utilized in conjunction with unpaired cells whenever available. It is straightforward to extend the GLUE framework to incorporate such pairing information, e.g., by adding another loss term that penalizes the embedding distances between paired cells[55]. Such an extension may ultimately lead to a solution for the general case of mosaic integration[14].

Apart from multi-omics integration, we also note that the GLUE framework could be suitable for cross-species integration, especially when distal species are concerned and one-to-one orthologs are limited. Specifically, we may compile all orthologs into a GLUE guidance graph and perform integration without explicit ortholog conversion. Under that setting, the GLUE approach could also be conceptually connected to a recent work called SAMap[56].

We believe that GLUE, as a modular and generalizable framework, creates an unprecedented opportunity towards effectively delineating gene regulatory maps via large-scale multi-omics integration at single-cell resolution. The whole package of GLUE, along with tutorials and demo cases, is available online at https://github.com/gao-lab/GLUE for the community.

Methods

The GLUE framework

We assume that there are K different omics layers to be integrated, each with a distinct feature set $\mathcal{V}_k, k = 1,2,\dots,K$. For example, in scRNA-seq, $\mathcal{V}_k$ is the set of genes, while in scATAC-seq, $\mathcal{V}_k$ is the set of chromatin regions. The data spaces of different omics layers are denoted as $\mathcal{X}_k \subseteq \mathbb{R}^{|\mathcal{V}_k|}$ with varying dimensionalities. We use $\mathbf{x}_k{}^{(n)} \in \mathcal{X}_k, n = 1,2,\dots,N_K$ to denote cells from the kth omics layer and $\mathbf{x}_{k_i}{}^{(n)}, i \in \mathcal{V}_k$ to denote the observed value of feature i of the kth layer in the nth cell. N_K is the sample size of the kth layer. Notably, the cells from different omics layers are unpaired and can have different sample sizes. To avoid cluttering, we drop the superscript (n) when referring to an arbitrary cell.

We model the observed data from different omics layers as generated by a low-dimensional latent variable (i.e., cell embedding) $\mathbf{u} \in \mathbb{R}^m$:

$$p(\mathbf{x}_k; \theta_k) = \int p(\mathbf{x}_k|\mathbf{u}; \theta_k) p(\mathbf{u}) d\mathbf{u} \qquad \text{Eq. 1}$$

where $p(\mathbf{u})$ is the prior distribution of the latent variable, $p(\mathbf{x}_k|\mathbf{u}; \theta_k)$ are learnable generative distributions (i.e., data decoders), and θ_k denotes learnable parameters in the decoders. The cell latent variable $\mathbf{u}$ is shared across different omics layers. In other words, $\mathbf{u}$ represents the common cell states underlying all omics observations, while the observed data from each layer are generated by a specific type of measurement of the underlying cell states.

With the introduction of variational posteriors $q(\mathbf{u}|\mathbf{x}_k; \phi_k)$ (i.e., data encoders, where ϕ_k are learnable parameters in the encoders), model fitting can be efficiently performed by maximizing the following evidence lower bounds:

$$\mathcal{L}_{\mathcal{X}_k}(\phi_k, \psi_k) = \mathbb{E}_{\mathbf{x}_k \sim p_{\text{data}}(\mathbf{x}_k)} \left[\mathbb{E}_{\mathbf{u} \sim q(\mathbf{u}|\mathbf{x}_k; \phi_k)} \log p(\mathbf{x}_k|\mathbf{u}; \theta_k) - \text{KL}\big(q(\mathbf{u}|\mathbf{x}_k; \phi_k) \parallel p(\mathbf{u})\big) \right] \qquad \text{Eq. 2}$$

Since different autoencoders are independently parameterized and trained on separate data, the cell embeddings learned for different omics layers could have inconsistent semantic meanings unless they are linked properly.

To link the autoencoders, we propose a guidance graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, which incorporates prior knowledge about the regulatory interactions among features at distinct omics layers, where $\mathcal{V} = \bigcup_{k=1}^{K} \mathcal{V}_k$ is the universal feature set and $\mathcal{E} = \{(i,j) | i,j \in \mathcal{V}\}$ is the set of edges. Each edge is also associated with signs and weights, which are denoted as s_{ij} and w_{ij}, respectively. We require that $w_{ij} \in (0,1]$, which can be interpreted as interaction credibility, and that $s_{ij} \in \{-1,1\}$, which specifies the sign of the regulatory interaction. For example, an ATAC peak located near the promoter of a gene is usually assumed to positively regulate its expression, so they can be connected with a positive edge ($s_{ij} = 1$). Meanwhile, DNA methylation in the gene promoter is usually assumed to suppress expression, so they can be connected with a negative edge ($s_{ij} = -1$). In addition to the connections between features, self-loops are also added for numerical stability, with $s_{ii} = 1, w_{ii} = 1, \forall i \in \mathcal{V}$.

We treat the guidance graph as observed variable and model it as generated by low-dimensional feature latent variables (i.e., feature embeddings) $\mathbf{v}_i \in \mathbb{R}^m, i \in \mathcal{V}$. Furthermore, differing from the previous model, we now model $\mathbf{x}_k$ as generated by the combination of feature latent variables $\mathbf{v}_i \in \mathbb{R}^m, i \in \mathcal{V}_k$ and the cell latent variable $\mathbf{u} \in \mathbb{R}^m$. For convenience, we introduce the notation $\mathbf{V} \in \mathbb{R}^{m \times |\mathcal{V}|}$, which combines all feature embeddings into a single matrix. The model likelihood can thus be written as:

$$p(\mathbf{x}_k, \mathcal{G}; \theta_k, \theta_{\mathcal{G}}) = \int p(\mathbf{x}_k | \mathbf{u}, \mathbf{V}; \theta_k) p(\mathcal{G} | \mathbf{V}; \theta_{\mathcal{G}}) p(\mathbf{u}) p(\mathbf{V}) d\mathbf{u} d\mathbf{V} \qquad \text{Eq. 3}$$

where $p(\mathbf{x}_k | \mathbf{u}, \mathbf{V}; \theta_k)$ and $p(\mathcal{G} | \mathbf{V}; \theta_{\mathcal{G}})$ are learnable generative distributions for the omics data (i.e., data decoders) and knowledge graph (i.e., graph decoder), respectively. θ_k and $\theta_{\mathcal{G}}$ are learnable parameters in the decoders. $p(\mathbf{u})$ and $p(\mathbf{V})$ are the prior distributions of the cell latent variable and feature latent variables, respectively, which are fixed as standard normal distributions for simplicity:

$$p(\mathbf{u}) = \mathcal{N}(\mathbf{u}; \mathbf{0}, \mathbf{I}_m) \qquad \text{Eq. 4}$$

$$p(\mathbf{v}_i) = \mathcal{N}(\mathbf{v}_i; \mathbf{0}, \mathbf{I}_m), \quad p(\mathbf{V}) = \prod_{i \in \mathcal{V}} p(\mathbf{v}_i) \qquad \text{Eq. 5}$$

though alternatives may also be used[57]. For convenience, we also introduce the notation $\mathbf{V}_k \in \mathbb{R}^{m \times |\mathcal{V}_k|}$, which contains only feature embeddings in the kth omics layer, and $\mathbf{u}_k$, which emphasizes that the cell embedding is from a cell in the kth omics layer.

The graph likelihood $p(\mathcal{G} | \mathbf{V}; \theta_{\mathcal{G}})$ (i.e., graph decoder) is defined as:

$$p(\mathcal{G}|\mathbf{V};\theta_{\mathcal{G}}) = \mathbb{E}_{i,j\sim p(i,j;w_{ij})}\left[\sigma(s_{ij}\cdot\mathbf{v}_i^{\top}\mathbf{v}_j)\cdot\mathbb{E}_{j'\sim p_{\mathrm{ns}}(j'|i)}\left(1-\sigma(s_{ij}\cdot\mathbf{v}_i^{\top}\mathbf{v}_{j'})\right)\right] \quad \text{Eq. 6}$$

where σ is the sigmoid function and p_{ns} is a negative sampling distribution[58]. In other words, we first sample the edges (i,j) with probabilities proportional to the edge weights and then sample vertices j' that are not connected to i and treat them as if $s_{ij'} = s_{ij}$. When maximizing the graph likelihood, the inner products between features are maximized or minimized (per edge sign) based on the Bernoulli distribution. For example, ATAC peaks located near the promoter of a gene would be encouraged to have similar embeddings to that of the gene, while DNA methylation in the gene promoter would be encouraged to have a dissimilar embedding to that of the gene.

The data likelihoods $p(\mathbf{x}_k|\mathbf{u},\mathbf{V};\theta_k)$ (i.e., data decoders) in Eq. 3 are built upon the inner product between the cell embedding $\mathbf{u}$ and feature embeddings $\mathbf{V}_k$. Thus, analogous to the loading matrix in principal component analysis (PCA), the feature embeddings $\mathbf{V}_k$ confer semantic meanings for the cell embedding space. As $\mathbf{V}_k$ are modulated by interactions among omics features in the guidance graph, the semantic meanings become linked. The exact formulation of data likelihood depends on the omics data distribution. For example, for count-based scRNA-seq and scATAC-seq data, we used the negative binomial (NB) distribution:

$$p(\mathbf{x}_k|\mathbf{u},\mathbf{V};\theta_k) = \prod_{i\in\mathcal{V}_k}\mathrm{NB}(\mathbf{x}_{k_i};\boldsymbol{\mu}_i,\boldsymbol{\theta}_i) \quad \text{Eq. 7}$$

$$\mathrm{NB}(\mathbf{x}_{k_i};\boldsymbol{\mu}_i,\boldsymbol{\theta}_i) = \frac{\Gamma(\mathbf{x}_{k_i}+\boldsymbol{\theta}_i)}{\Gamma(\boldsymbol{\theta}_i)\Gamma(\mathbf{x}_{k_i}+1)}\left(\frac{\boldsymbol{\mu}_i}{\boldsymbol{\theta}_i+\boldsymbol{\mu}_i}\right)^{\mathbf{x}_{k_i}}\left(\frac{\boldsymbol{\theta}_i}{\boldsymbol{\theta}_i+\boldsymbol{\mu}_i}\right)^{\boldsymbol{\theta}_i} \quad \text{Eq. 8}$$

$$\boldsymbol{\mu}_i = \mathrm{Softmax}_i(\boldsymbol{\alpha}\odot\mathbf{V}_k^{\top}\boldsymbol{u}+\boldsymbol{\beta})\cdot\sum_{j\in\mathcal{V}_k}\mathbf{x}_{k_j} \quad \text{Eq. 9}$$

where $\boldsymbol{\mu},\boldsymbol{\theta}\in\mathbb{R}_+^{|\mathcal{V}_k|}$ are the mean and dispersion of the NB distribution, respectively, and $\boldsymbol{\alpha}\in\mathbb{R}_+^{|\mathcal{V}_k|}$, $\boldsymbol{\beta}\in\mathbb{R}^{|\mathcal{V}_k|}$ are scaling and bias factors. $\odot$ is the Hadamard product. $\mathrm{Softmax}_i$ represents the ith dimension of the softmax output. The set of learnable parameters is $\theta_k = \{\boldsymbol{\theta},\boldsymbol{\alpha},\boldsymbol{\beta}\}$. Analogously, many other distributions can also be supported, as long as we can parameterize the means of the distributions by feature-cell inner products.

For efficient inference and optimization, we introduce the following factorized variational posterior:

$$q(\mathbf{u},\mathbf{V}|\mathbf{x}_k,\mathcal{G};\phi_k,\phi_{\mathcal{G}}) = q(\mathbf{u}|\mathbf{x}_k;\phi_k)\cdot q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}}) \quad \text{Eq. 10}$$

The graph variational posterior $q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}})$ (i.e., graph encoder) is modeled as diagonal-covariance normal distributions parameterized by a graph convolutional network[59]:

$$q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}}) = \prod_{i\in\mathcal{V}} q(\mathbf{v}_i|\mathcal{G};\phi_{\mathcal{G}}) \qquad \text{Eq. 11}$$

$$q(\mathbf{v}_i|\mathcal{G};\phi_{\mathcal{G}}) = \mathcal{N}\left(\mathbf{v}_i; \mathrm{GCN}_{\boldsymbol{\mu}_i}(\mathcal{G};\phi_{\mathcal{G}}), \mathrm{GCN}_{\boldsymbol{\sigma}_i^2}(\mathcal{G};\phi_{\mathcal{G}})\right) \qquad \text{Eq. 12}$$

where $\phi_{\mathcal{G}}$ represents the learnable parameters in the GCN encoder.

The variational data posteriors $q(\mathbf{u}|\mathbf{x}_k;\phi_k)$ (i.e., data encoders) are modeled as diagonal-covariance normal distributions parameterized by multilayer perceptron (MLP) neural networks:

$$q(\mathbf{u}|\mathbf{x}_k,\mathbf{V}_k;\phi_k) = \mathcal{N}\left(\mathbf{u}; \mathrm{MLP}_{k,\boldsymbol{\mu}}(\mathbf{x}_k;\phi_k), \mathrm{MLP}_{k,\boldsymbol{\sigma}^2}(\mathbf{x}_k;\phi_k)\right) \qquad \text{Eq. 13}$$

where ϕ_k is the set of learnable parameters in the MLP encoder of the kth omics layer.

Model fitting can then be performed by maximizing the following evidence lower bound:

$$\sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}_k\sim p_{\text{data}}(\mathbf{x}_k)} \begin{bmatrix} \mathbb{E}_{\mathbf{u}\sim q(\mathbf{u}|\mathbf{x}_k;\phi_k),\mathbf{V}\sim q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}})} \log p(\mathbf{x}_k|\mathbf{u},\mathbf{V};\theta_k) p(\mathcal{G}|\mathbf{V};\theta_{\mathcal{G}}) \\ -\mathrm{KL}\left(q(\mathbf{u}|\mathbf{x}_k;\phi_k) q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}}) \parallel p(\mathbf{u})p(\mathbf{V})\right) \end{bmatrix} \qquad \text{Eq. 14}$$

which can be further rearranged into the following form:

$$K \cdot \mathcal{L}_{\mathcal{G}}(\theta_{\mathcal{G}},\phi_{\mathcal{G}}) + \sum_{k=1}^{K} \mathcal{L}_{\mathcal{X}_k}(\theta_k,\phi_k,\phi_{\mathcal{G}}) \qquad \text{Eq. 15}$$

where we have

$$\mathcal{L}_{\mathcal{X}_k}(\theta_k,\phi_k,\phi_{\mathcal{G}}) = \mathbb{E}_{\mathbf{x}_k\sim p_{\text{data}}(\mathbf{x}_k)} \left[\mathbb{E}_{\mathbf{u}\sim q(\mathbf{u}|\mathbf{x}_k;\phi_k),\mathbf{V}\sim q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}})} \log p(\mathbf{x}_k|\mathbf{u},\mathbf{V};\theta_k) - \mathrm{KL}\big(q(\mathbf{u}|\mathbf{x}_k;\phi_k) \parallel p(\mathbf{u})\big)\right] \qquad \text{Eq. 16}$$

$$\mathcal{L}_{\mathcal{G}}(\theta_{\mathcal{G}},\phi_{\mathcal{G}}) = \mathbb{E}_{\mathbf{V}\sim q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}})} \log p(\mathcal{G}|\mathbf{V};\theta_{\mathcal{G}}) - \mathrm{KL}\left(q(\mathbf{V}|\mathcal{G};\phi_{\mathcal{G}}) \parallel p(\mathbf{V})\right) \qquad \text{Eq. 17}$$

Below, for convenience, we denote the union of all encoder parameters as $\phi = (\bigcup_{k=1}^{K} \phi_k) \cup \phi_{\mathcal{G}}$ and the union of all decoder parameters as $\theta = (\bigcup_{k=1}^{K} \theta_k) \cup \theta_{\mathcal{G}}$.

To ensure the proper alignment of different omics layers, we use the adversarial alignment strategy[27, 60]. A discriminator D with a K-dimensional softmax output is introduced, which predicts the omics layers of cells based on their embeddings $\mathbf{u}$. The discriminator D is trained by minimizing the multiclass classification cross entropy:

$$\mathcal{L}_{\mathrm{D}}(\phi,\psi) = -\frac{1}{K}\sum_{k=1}^{K} \mathbb{E}_{\mathbf{x}_k\sim p_{\text{data}}(\mathbf{x}_k)} \mathbb{E}_{\mathbf{u}\sim q(\mathbf{u}|\mathbf{x}_k;\phi_k)} \log \mathrm{D}_k(\mathbf{u};\psi) \qquad \text{Eq. 18}$$

where D_k represents the kth dimension of the discriminator output and ψ is the set of learnable parameters in the discriminator. The data encoders can then be trained in the opposite direction to

fool the discriminator, ultimately leading to the alignment of cell embeddings from different omics layers[61].

The overall training objective of GLUE thus consists of:

$$\min_{\psi} \lambda_{\mathrm{D}} \cdot \mathcal{L}_{\mathrm{D}}(\phi, \psi) \qquad \text{Eq. 19}$$

$$\max_{\theta,\phi} \lambda_{\mathrm{D}} \cdot \mathcal{L}_{\mathrm{D}}(\phi, \psi) + \lambda_{\mathcal{G}} K \cdot \mathcal{L}_{\mathcal{G}}(\theta_{\mathcal{G}}, \phi_{\mathcal{G}}) + \sum_{k=1}^{K} \mathcal{L}_{\mathcal{X}_k}(\theta_k, \phi_k, \phi_{\mathcal{G}}) \qquad \text{Eq. 20}$$

The two hyperparameters λ_{D} and $\lambda_{\mathcal{G}}$ control the contributions of adversarial alignment and graph-based feature embedding, respectively. We use stochastic gradient descent (SGD) to train the GLUE model. Each SGD iteration is divided into two steps. In the first step, the discriminator is updated according to objective Eq. 19. In the second step, the data and graph autoencoders are updated according to Eq. 20. The RMSprop optimizer with no momentum term is employed to ensure the stability of adversarial training.

Implementation details

We applied linear dimensionality reduction using canonical methods such as PCA (for scRNA-seq) or LSI (latent semantic indexing, for scATAC-seq) as the first transformation layers of the data encoders (note that the decoders were still fitted in the original feature spaces). This effectively reduced model size and enabled a modular input, so advanced dimensionality reduction or batch effect correction methods can also be used instead as preprocessing steps for GLUE integration.

To ensure stable alignment, we used batch normalization in the data encoder layers and employed additive noise annealing. Specifically, noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{\epsilon}; \mathbf{0}, \tau \cdot \mathbf{I}_m)$ was added to the cell embeddings $\mathbf{u}$ before passing to the discriminator. The parameter τ controls the noise level, which starts at $\tau = 1$ and decreases linearly per epoch until reaching 0 (i.e., noise annealing). The number of annealing epochs was set automatically based on the data size and learning rate to match a learning progress equivalent to 4,000 iterations at a learning rate of 0.002.

During model training, 10% of the cells were used as the validation set. In the final stage of training, the learning rate would be reduced by factors of 10 if the validation loss did not improve for consecutive epochs. Training would be terminated if the validation loss still did not improve for consecutive epochs. The patience for learning rate reduction, training termination, and the maximal

number of training epochs were automatically set based on the data size and learning rate to match a learning progress equivalent to 1,000, 2,000, and 16,000 iterations at a learning rate of 0.002, respectively.

For all benchmarks and case studies with GLUE, we used the default hyperparameters unless explicitly stated. The set of default hyperparameters is presented in Supplementary Fig. 4.

Systematic benchmarks

UnionCom[20] and GLUE were executed using the Python packages "unioncom" (v0.3.0) and "scglue" (v0.1.0), respectively. LIGER[16], bindSC[29], and Seurat v3[15] were executed using the R packages "rliger" (v0.5.0), "bindSC" (v1.0.0), and "Seurat" (v4.0.2), respectively. For each method, we used the default hyperparameter settings and data preprocessing steps as recommended. For the scRNA-seq data, 2,000 highly variable genes were selected using the Seurat "vst" method. To construct the guidance graph, we connected ATAC peaks with RNA genes via positive edges if they overlapped in either the gene body or proximal promoter regions (defined as 2 kb upstream from the TSS). For the methods that require feature conversion (LIGER, bindSC, and Seurat v3), we converted the scATAC-seq data to gene-level activity scores by summing up counts in the ATAC peaks connected to specific genes in the guidance graph.

MAP (mean average precision) was used to evaluate the cell type resolution. Supposing that the cell type of the ith cell is $y^{(i)}$ and that the cell types of its K ordered nearest neighbors are $y_1^{(i)}, y_2^{(i)}, \ldots y_K^{(i)}$, the MAP is then defined as follows:

$$\mathrm{MAP} = \frac{1}{N}\sum_{i=1}^{N} \mathrm{AP}^{(i)} \qquad \text{Eq. 21}$$

$$\mathrm{AP}^{(i)} = \begin{cases} \dfrac{\sum_{k=1}^{K} \mathbb{1}_{y^{(i)}=y_k} \cdot \dfrac{\sum_{j=1}^{k} \mathbb{1}_{y^{(i)}=y_j^{(i)}}}{k}}{\sum_{k=1}^{K} \mathbb{1}_{y^{(i)}=y_k^{(i)}}}, & \text{if } \sum_{k=1}^{K} \mathbb{1}_{y^{(i)}=y_k^{(i)}} > 0 \\ 0, & \text{otherwise} \end{cases} \qquad \text{Eq. 22}$$

where $\mathbb{1}_{y^{(i)}=y_k^{(i)}}$ is an indicator function that equals 1 if $y^{(i)} = y_k^{(i)}$ and 0 otherwise. For each cell, AP (average precision) computes the average cell type precision up to each cell type-matched neighbor, and MAP is the average AP across all cells. We set K to 1% of the total number of cells in each dataset. MAP has a range of $0 - 1$, and higher values indicate better cell type resolution.

SAS (Seurat alignment score) was used to evaluate the extent of mixing among distinct omics layers and was computed as described in the original paper[31]:

$$\mathrm{SAS} = 1 - \frac{\bar{x} - \frac{k}{N}}{k - \frac{k}{N}} \qquad \text{Eq. 23}$$

where $\bar{x}$ is the average number of cells from the same omics layer among the k nearest neighbors (different layers were first subsampled to the same size), and N is the number of omics layers. SAS has a range of $0 - 1$, and higher values indicate better mixing.

FOSCTTM (fraction of samples closer than the true match)[32] was used to evaluate the single-cell level alignment accuracy. It was computed on two datasets with known cell-to-cell pairings. Suppose that each dataset contains N cells, and that the cells are sorted in the same order, i.e., the ith cell in the first dataset is paired with the ith cell in the second dataset. Denote $\mathbf{x}$ and $\mathbf{y}$ as the cell embeddings of the first and second dataset, respectively. The FOSCTTM is then defined as:

$$\mathrm{FOSCTTM} = \frac{1}{2N}\left(\sum_{i=1}^{N} \frac{n_1^{(i)}}{N} + \sum_{i=1}^{N} \frac{n_2^{(i)}}{N}\right) \qquad \text{Eq. 24}$$

$$n_1^{(i)} = \left|\left\{j \middle| d\left(\mathbf{x}_j, \mathbf{y}_i\right) < d\left(\mathbf{x}_i, \mathbf{y}_i\right)\right\}\right| \qquad \text{Eq. 25}$$

$$n_2^{(i)} = \left|\left\{j \middle| d\left(\mathbf{x}_i, \mathbf{y}_j\right) < d\left(\mathbf{x}_i, \mathbf{y}_i\right)\right\}\right| \qquad \text{Eq. 26}$$

where $n_1^{(i)}$ and $n_2^{(i)}$ are the number of cells in the first and second dataset, respectively, that are closer to the ith cell than their true matches in the opposite dataset. d is the Euclidean distance.

For the baseline benchmark, each method was run 8 times with different random seeds, except for bindSC, which has a deterministic implementation and was run only once. For the guidance corruption benchmark, we removed the specified proportions of existing peak-gene interactions and added equal numbers of nonexistent interactions, so the total number of interactions remained unchanged. Of note, feature conversion was also repeated using the corrupted guidance graphs. The corruption procedure was repeated 8 times with different random seeds. For the subsampling benchmark, the scRNA-seq and scATAC-seq cells were subsampled in pairs (so FOSCTTM could still be computed). The subsampling process was also repeated 8 times with different random seeds.

For the systematic scalability test (Supplementary Fig. 13a), all methods were run on a Linux workstation with 40 CPU cores (two Intel Xeon Silver 4210 chips), 250 GB of RAM, and NVIDIA GeForce RTX 2080 Ti GPUs. Only a single GPU card was used when training GLUE.

Triple-omics integration

The scRNA-seq and scATAC-seq data were handled as previously described (see the section "Systematic benchmarks"). Due to low coverage per single-C site, the snmC-seq data were converted to average methylation levels in gene bodies. The mCH and mCG levels were quantified separately, resulting in 2 features per gene. The gene methylation levels were normalized by the global methylation level per cell. An initial dimensionality reduction was performed using PCA (see the section "Implementation details"). For the triple-omics guidance graph, the mCH and mCG levels were connected to the corresponding genes with negative edges.

The normalized methylation levels were positive, with dropouts corresponding to the genes that were not covered in single cells. As such, we used the zero-inflated log-normal (ZILN) distribution for the data decoder:

$$p(\mathbf{x}_k|\mathbf{u},\mathbf{V};\theta_k) = \prod_{i\in\mathcal{V}_k} \mathrm{ZILN}\left(\mathbf{x}_{k_i};\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i,\boldsymbol{\delta}_i\right) \qquad \text{Eq. 27}$$

$$\mathrm{ZILN}\left(\mathbf{x}_{k_i};\boldsymbol{\mu}_i,\boldsymbol{\sigma}_i,\boldsymbol{\delta}_i\right) = \begin{cases} \dfrac{1-\boldsymbol{\delta}_i}{\mathbf{x}_{k_i}\boldsymbol{\sigma}_i\sqrt{2\pi}}\exp\left(-\dfrac{\left(\log\mathbf{x}_{k_i}-\boldsymbol{\mu}_i\right)^2}{2{\boldsymbol{\sigma}_i}^2}\right), & \mathbf{x}_{k_i}>0 \\ \boldsymbol{\delta}_i, & \mathbf{x}_{k_i}=0 \end{cases} \qquad \text{Eq. 28}$$

$$\boldsymbol{\mu}_i = \boldsymbol{\alpha}\odot\mathbf{V}_k^{\top}\boldsymbol{u}+\boldsymbol{\beta} \qquad \text{Eq. 29}$$

where $\boldsymbol{\mu}\in\mathbb{R}^{|\mathcal{V}_k|}, \boldsymbol{\sigma}\in\mathbb{R}_+^{|\mathcal{V}_k|}, \boldsymbol{\delta}\in(0,1)^{|\mathcal{V}_k|}$ are the log-scale mean, log-scale standard deviation and zero-inflation parameters of the ZILN distribution, respectively, and $\boldsymbol{\alpha}\in\mathbb{R}_+^{|\mathcal{V}_k|}, \boldsymbol{\beta}\in\mathbb{R}^{|\mathcal{V}_k|}$ are scaling and bias factors.

To unify the cell type labels, we performed a nearest neighbor-based label transfer with the snmC-seq dataset as a reference. The 5 nearest neighbors in snmC-seq were identified for each scRNA-seq and scATAC-seq cell in the aligned embedding space, and majority voting was used to determine the transferred label. To verify whether the alignment was correct, we tested for significant overlap in cell type marker genes. The features of all omics layers were first converted to genes. Then, for each omics layer, the cell type markers were identified using the one-vs.-rest Wilcoxon rank-sum test with

the following criteria: FDR < 0.05 and log-fold change > 0 for scRNA-seq/scATAC-seq; FDR < 0.05 and log-fold change < 0 for snmC-seq. The significance of marker overlap was determined by the three-way Fisher's exact test[36].

To perform correlation and regression analysis after the integration, we clustered all cells from the three omics layers using fine-scale k-means (k = 200). Then, for each omics layer, the cells in each cluster were aggregated into a metacell by summing their expression/accessibility counts or averaging their DNA methylation levels. Therefore, the metacells were established as paired samples, based on which feature correlation and regression analyses could be conducted.

Model-based cis-regulatory inference

To ensure consistency of cell types, we first selected the overlapping cell types between the 10x Multiome and pcHi-C data. The remaining cell types included T cells, B cells and monocytes. The eQTL data were used as is, because they were not cell type-specific. For scRNA-seq, we selected 6,000 highly variable genes. For the initial regulatory inference, the guidance graph was constructed by connecting RNA genes with ATAC peaks within 150 kb of the gene promoters (defined as 2 kb upstream from the TSS); the graph was weighted by a power-law function $w = (d + 1)^{-0.75}$ (d is the genomic distance in kb), which has been proposed to model the probability of chromatin contact[38, 39].

To incorporate the regulatory evidence of pcHi-C and eQTL, we anchored all evidence to that between the ATAC peaks and RNA genes. A peak-gene pair was considered supported by pcHi-C if (1) the gene promoter was within 1 kb of a bait fragment, (2) the peak was within 1 kb of an other-end fragment, and (3) significant contact was identified between the bait and the other-end fragment in pcHi-C. The pcHi-C-supported peak-gene interactions were weighted by multiplying the promoter-to-bait and the peak-to-other-end power-law weights (see above). If a peak-gene pair was supported by multiple pcHi-C contacts, the weights were summed and clipped at a maximum of 1. A peak-gene pair was considered supported by eQTL if (1) the peak overlapped an eQTL locus and (2) the locus was associated with the expression of the gene. The eQTL-supported peak-gene interactions were assigned weights of 1. The composite guidance graph was constructed by adding the pcHi-C- and eQTL-supported interactions to the previous distance-based interactions, allowing for multi-edges.

For regulatory inference, only peak-gene pairs within 150 kb in distance were considered. The GLUE training process was repeated 4 times with different random seeds. For each repeat, the peak-gene regulatory score was computed as the cosine similarity between the feature embeddings. The final regulatory inference was obtained by averaging the regulatory scores across the 4 repeats.

TF-target gene regulatory inference

We employed the SCENIC workflow[62] to construct a TF-gene regulatory network from the inferred peak-gene regulatory interactions. Briefly, the SCENIC workflow first constructs a gene coexpression network based on the scRNA-seq data, and then uses external cis-regulatory evidence to filter out false positives. SCENIC accepts cis-regulatory evidence in the form of gene rankings per TF, i.e., genes with higher TF enrichment levels in their regulatory regions are ranked higher. To construct the rankings based on our inferred peak-gene interactions, we first overlapped the ENCODE TF ChIP peaks[63] with the ATAC peaks and counted the number of ChIP peaks for each TF in each ATAC peak. Since different genes can have different numbers of connected ATAC peaks, and the ATAC peaks vary in length (longer peaks can contain more ChIP peaks by chance), we devised a sampling-based approach to evaluate TF enrichment. Specifically, for each gene, we randomly sampled 1,000 sets of ATAC peaks that matched the connected ATAC peaks in both number and length distribution. We counted the numbers of TF ChIP peaks in these random ATAC peaks as null distributions. For each TF in each gene, an empirical *P* value could then be computed by comparing the observed number of ChIP peaks to the null distribution. Finally, we ranked the genes by the empirical *P* values for each TF, producing the cis-regulatory rankings used by SCENIC. Since peak-gene-based inference is mainly focused on remote regulatory regions, proximal promoters could be missed. As such, we provided SCENIC with both the above peak-based and proximal promoter-based cis-regulatory rankings.

Integration for the human multi-omics atlas

The scRNA-seq and scATAC-seq atlases have highly unbalanced cell type compositions, which is primarily caused by differences in organ sampling sizes (Supplementary Fig. 13b). Although cell types are unknown during real-world analyses, organ sources are typically available and can be utilized to help balance the integration process. To perform organ-balanced data preprocessing, we first subsampled each omics layer to match the organ compositions. For the scRNA-seq data, 4,000 highly variable genes were selected using the organ-balanced subsample. Then, for the initial

dimensionality reduction, we fitted PCA (scRNA-seq) and LSI (scATAC-seq) on the organ-balanced subsample and applied the projection to the full data. The PCA/LSI coordinates were used as the first transformation layer in the GLUE data encoders (see the section "Implementation details"), as well as for metacell aggregation (see below). The guidance graph was constructed as described previously (see the section "Systematic benchmarks").

The two atlases consist of large numbers of cells but with low coverage per cell. To alleviate dropout and increase the training speed simultaneously, we designed a multistage training strategy, where the GLUE model was pretrained on aggregated metacells and then fine-tuned on the original single cells. Specifically, in the first stage, we clustered the cells in each omics layer using fine-scaled k-means (k = 100,000 for scRNA-seq and k = 40,000 for scATAC-seq). To balance the organ compositions at the same time, k-means centroids were fitted on the previous organ-balanced subsample and then applied to the full data. The cells in each k-means cluster were aggregated into a metacell by summing their expression/accessibility counts and averaging their PCA/LSI coordinates. GLUE was then pretrained on the aggregated metacells without adversarial alignment, which roughly oriented the cell embeddings but did not actually align them. To better utilize the large data size, the hidden layer dimensionality was doubled to 512 from the default 256.

In the second stage, GLUE was fine-tuned on the full single-cell data with weighted adversarial alignment (see below). As shown in previous work[27], pure adversarial alignment amounts to minimizing a generalized form of Jensen–Shannon divergence among the cell embedding distributions of different omics layers:

$$\frac{1}{K}\sum_{k=1}^{K}\mathrm{KL}\left(q_k(\mathbf{u})\middle\|\frac{1}{K}\sum_{k=1}^{K}q_k(\mathbf{u})\right) \qquad \text{Eq. 30}$$

where $q_k(\mathbf{u}) = \mathbb{E}_{\mathbf{x}_k\sim p_{\mathrm{data}}(\mathbf{x}_k)}q(\mathbf{u}|\mathbf{x}_k;\phi_k)$ represents the marginal cell embedding distribution of the kth layer. Without other loss terms, Eq. 30 converges at perfect alignment, i.e., when $q_i(\mathbf{u}) = q_j(\mathbf{u}), \forall i \neq j$. This can be problematic when cell type compositions differ dramatically across different layers, as was the case here. To address this issue, we added cell-specific weights $w^{(n)}$ to the discriminator loss in Eq. 18:

$$\mathcal{L}_{\mathrm{D}}(\phi,\psi) = -\frac{1}{K}\sum_{k=1}^{K}\frac{1}{W_k}\sum_{n=1}^{N_k}w^{(n)}\cdot\mathbb{E}_{\mathbf{u}\sim q\left(\mathbf{u}|\mathbf{x}_k^{(n)};\phi_k\right)}\log \mathrm{D}_k(\mathbf{u};\psi) \qquad \text{Eq. 31}$$

where the normalizer $W_k = \sum_{n=1}^{N_k} w^{(n)}$. The adversarial alignment still amounts to minimizing Eq. 30 but with weighted marginal cell embedding distributions $q_k(\mathbf{u}) = \frac{1}{W_k}\sum_{n=1}^{N_k} w^{(n)} q\left(\mathbf{u}|\mathbf{x}_k^{(n)}; \phi_k\right)$. By assigning appropriate weights to balance the cell distributions across different layers, the optimum of $q_i(\mathbf{u}) = q_j(\mathbf{u}), \forall i \neq j$ could be much closer to the desired alignment. For example, a viable choice for the cell weights is organ-balancing weights. Suppose that the organ proportions in scRNA-seq and scATAC-seq are $f_1, f_2, \dots, f_O$ and $g_1, g_2, \dots, g_O$ (O is the number of organs, $\sum_{i=1}^{O} f_i = 1$, $\sum_{i=1}^{O} g_i = 1$), respectively. We can weight the RNA cells in the ith organ by $\sqrt{g_i/f_i}$ and weight the ATAC cells in the ith organ by $\sqrt{f_i/g_i}$, so that the ith organ has a balanced accumulative contribution of $\sqrt{f_i g_i}$, regardless of omics layers. However, in case there are major differences among the cell type compositions within the same organ, organ-level balancing can be insufficient. As such, we designed a method to compute cell type-level balancing weights in an unsupervised manner. For each omics layer, we clustered the cell embeddings using Louvain clustering and matched the clusters in different layers via cosine similarity. The population size of each cluster was distributed to its matched counterparts by cosine similarity. Subsequently, for each omics layer, we could obtain the proportion of each cluster f_i and its effective proportion g_i in the opposite layer (by normalizing the effective population size received from the opposite layer). Balancing weights could then be computed as above. The weight-balanced alignment proved effective in aligning the highly skewed data (Fig. 5).

For a comparison with other integration methods, we also tried Seurat v3, which was the second-best method in our previous benchmark in terms of accuracy. Since Seurat v3 could not scale to the full dataset (Supplementary Fig. 13a), we only managed to apply it to the aggregated data, as used in the first stage of GLUE training, but the result was suboptimal (Supplementary Fig. 18).

Data availability

All datasets used in this study are already published and were obtained from public data repositories. See Supplementary Table 1 for detailed information, including access codes and URLs. All benchmark data are available in Supplementary Table 2.

Code availability

The GLUE framework was implemented in the "scglue" Python package, which is available at https://github.com/gao-lab/GLUE. For reproducibility, the scripts for all benchmarks and case studies were assembled using Snakemake, which is also available in the above repository.

Acknowledgments

The authors would like to thank Drs. Zemin Zhang, Xiaoliang Sunney Xie, Letian Tao, Cheng Li, Jian Lu (at Peking University) and Yang Ding (at the Beijing Institute of Radiation Medicine) for their helpful discussions and comments during the study, as well as authors of the datasets used in this work for their kindly help.

This work was supported by funds from the National Key Research and Development Program (2016YFC0901603), the China 863 Program (2015AA020108), and the State Key Laboratory of Protein and Plant Gene Research and the Beijing Advanced Innovation Center for Genomics (ICG) at Peking University. The research of G.G. was supported in part by the National Program for Support of Top-notch Young Professionals.

Part of the analysis was performed on the Computing Platform of the Center for Life Sciences of Peking University and supported by the High-performance Computing Platform of Peking University.

Author contributions

G.G. conceived the study and supervised the research; Z.J.C. designed and implemented the computational framework and conducted benchmarks and case studies, with guidance from G.G.; Z.J.C. and G.G. wrote the manuscript.

Competing interests

The authors declare no competing interests.

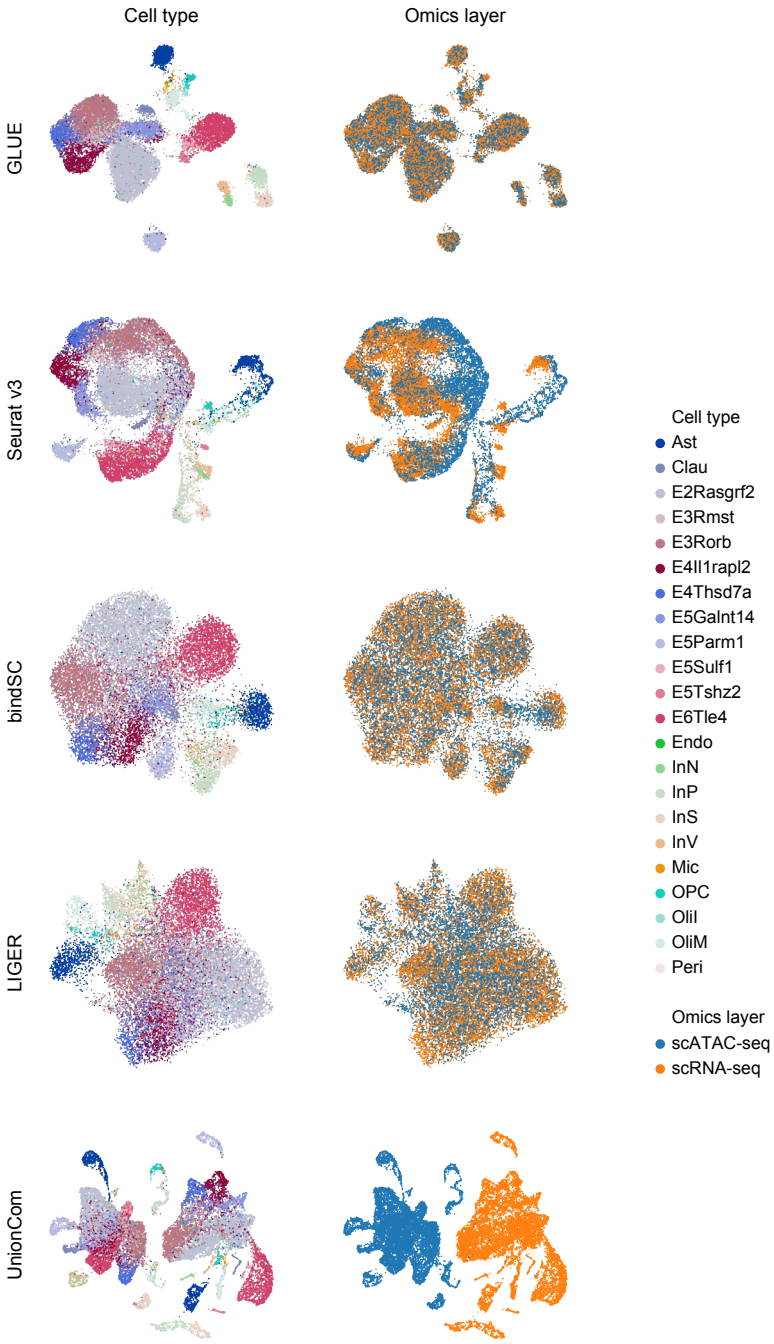
References

1. Cusanovich, D.A. et al. Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing. *Science* **348**, 910–914 (2015).
2. Chen, X., Miragaia, R.J., Natarajan, K.N. & Teichmann, S.A. A rapid and robust method for single cell chromatin accessibility profiling. *Nat. Commun.* **9**, 5345 (2018).
3. Luo, C. et al. Single-cell methylomes identify neuronal subtypes and regulatory elements in mammalian cortex. *Science* **357**, 600–604 (2017).
4. Mulqueen, R.M. et al. Highly scalable generation of DNA methylation profiles in single cells. *Nat. Biotechnol.* **36**, 428–431 (2018).
5. Picelli, S. et al. Smart-seq2 for sensitive full-length transcriptome profiling in single cells. *Nat. Methods* **10**, 1096–1098 (2013).
6. Zheng, G.X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
7. Packer, J. & Trapnell, C. Single-cell multi-omics: An engine for new quantitative models of gene regulation. *Trends Genet.* **34**, 653–665 (2018).
8. Chen, S., Lake, B.B. & Zhang, K. High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell. *Nat. Biotechnol.* **37**, 1452–1457 (2019).
9. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116 (2020).
10. Clark, S.J. et al. scNMT-seq enables joint profiling of chromatin accessibility DNA methylation and transcription in single cells. *Nat. Commun.* **9**, 781 (2018).
11. Wang, Y. et al. Single-cell multiomics sequencing reveals the functional regulatory landscape of early embryos. *Nat. Commun.* **12**, 1247 (2021).
12. Lake, B.B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
13. Bravo Gonzalez-Blas, C. et al. Identification of genomic enhancers through spatial integration of single-cell transcriptomics and epigenomics. *Mol. Syst. Biol.* **16**, e9438 (2020).
14. Argelaguet, R., Cuomo, A.S.E., Stegle, O. & Marioni, J.C. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.* (2021).
15. Stuart, T. et al. Comprehensive integration of single-cell data. *Cell* **177**, 1888–1902 (2019).
16. Welch, J.D. et al. Single-cell multi-omic integration compares and contrasts features of brain cell identity. *Cell* **177**, 1873–1887 (2019).
17. Chen, H. et al. Assessment of computational methods for the analysis of single-cell ATAC-seq data. *Genome Biol.* **20**, 241 (2019).
18. Duren, Z. 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).
19. Zeng, W. et al. DC3 is a method for deconvolution and coupled clustering from bulk and single-cell genomics data. *Nat. Commun.* **10**, 4613 (2019).
20. Cao, K., Bai, X., Hong, Y. & Wan, L. Unsupervised topological alignment for single-cell multi-omics integration. *Bioinformatics* **36**, i48–i56 (2020).
21. Demetci, P., Santorella, R., Sandstede, B., Noble, W.S. & Singh, R. Gromov-Wasserstein optimal transport to align single-cell multi-omics data. Preprint at https://www.biorxiv.org/content/10.1101/2020.04.28.066787 (2020).
22. Svensson, V., Vento-Tormo, R. & Teichmann, S.A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).

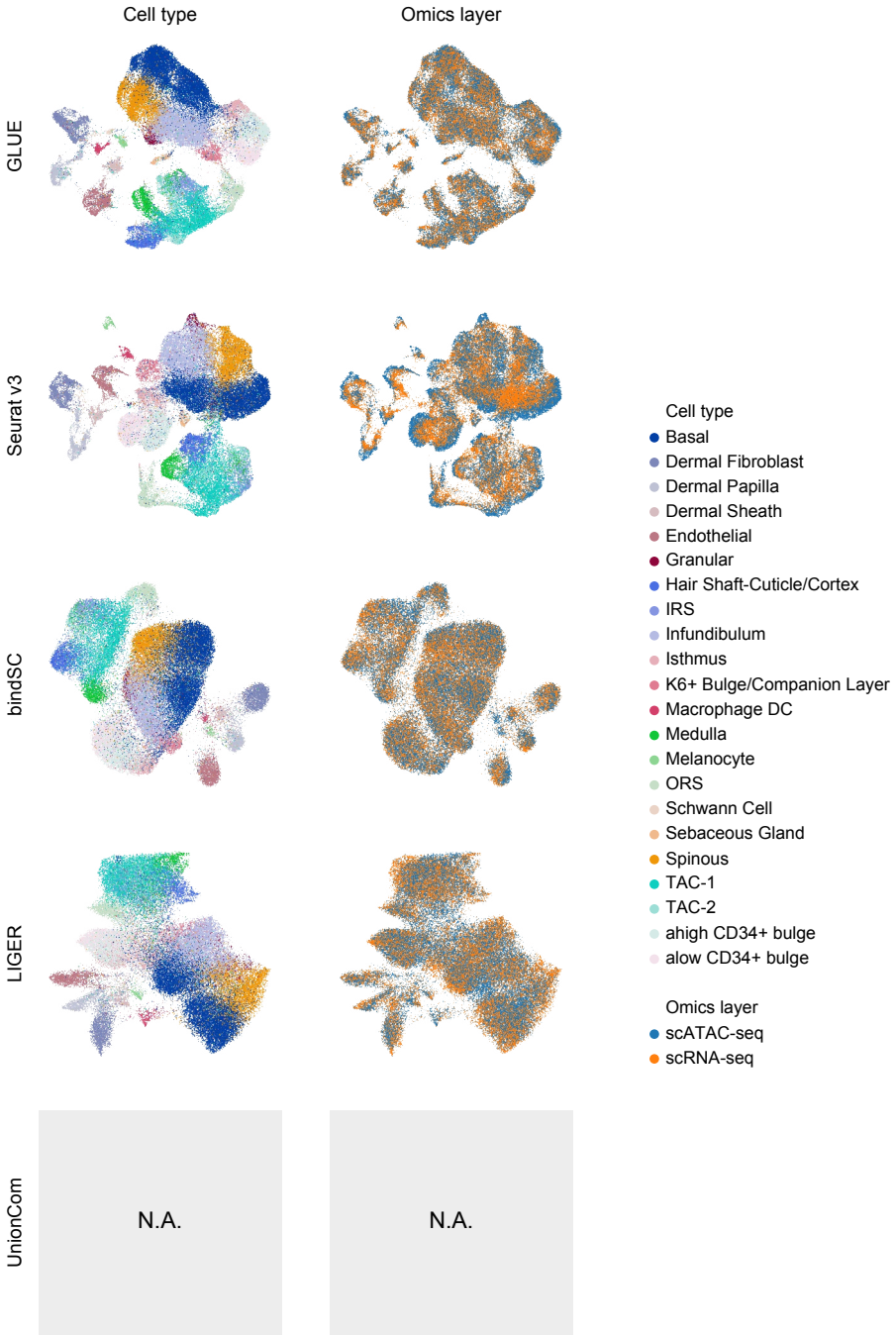
23. Kozareva, V. et al. A transcriptomic atlas of the mouse cerebellum reveals regional specializations and novel cell types. Preprint at http://biorxiv.org/content/early/2020/03/05/2020.03.04.976407.abstract (2020).
24. Cao, J. et al. A human cell atlas of fetal gene expression. *Science* **370**, eaba7721 (2020).
25. Domcke, S. et al. A human cell atlas of fetal chromatin accessibility. *Science* **370**, eaba7612 (2020).
26. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat. Methods* **15**, 1053–1058 (2018).
27. Cao, Z.J., Wei, L., Lu, S., Yang, D.C. & Gao, G. Searching large-scale scRNA-seq databases via unbiased cell embedding with Cell BLAST. *Nat. Commun.* **11**, 3458 (2020).
28. Kipf, T.N. & Welling, M. Variational graph auto-encoders. Preprint at https://arxiv.org/abs/1611.07308 (2016).
29. Dou, J. et al. Unbiased integration of single cell multi-omics data. Preprint at https://www.biorxiv.org/content/10.1101/2020.12.11.422014 (2020).
30. 10x Genomics. PBMC from a healthy donor, single cell multiome ATAC gene expression demonstration data by Cell Ranger ARC 1.0.0. https://support.10xgenomics.com/single-cell-multiome-atac-gex/datasets/1.0.0/pbmc_granulocyte_sorted_10k (2020).
31. Butler, A., Hoffman, P., Smibert, P., Papalexi, E. & Satija, R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nat. Biotechnol.* **36**, 411–420 (2018).
32. Singh, R. et al. Unsupervised manifold alignment for single-cell multi-omics data. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. (ACM, Virtual Event, USA, 2020).
33. Saunders, A. et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell* **174**, 1015–1030 (2018).
34. 10x Genomics. Fresh cortex from adult mouse brain (v1), single cell ATAC demonstration data by Cell Ranger 1.1.0. https://support.10xgenomics.com/single-cell-atac/datasets/1.1.0/atac_v1_adult_brain_fresh_5k (2019).
35. Mo, A. et al. Epigenomic signatures of neuronal diversity in the mammalian brain. *Neuron* **86**, 1369–1384 (2015).
36. Wang, M., Zhao, Y. & Zhang, B. Efficient test and visualization of multi-set intersections. *Sci. Rep.* **5**, 16923 (2015).
37. Gabel, H.W. et al. Disruption of DNA-methylation-dependent long gene repression in Rett syndrome. *Nature* **522**, 89–93 (2015).
38. Dekker, J., Marti-Renom, M.A. & Mirny, L.A. Exploring the three-dimensional organization of genomes: Interpreting chromatin interaction data. *Nat. Rev. Genet.* **14**, 390–403 (2013).
39. Pliner, H.A. et al. Cicero predicts cis-regulatory DNA interactions from single-cell chromatin accessibility data. *Mol. Cell* **71**, 858–871 (2018).
40. Javierre, B.M. et al. Lineage-specific genome architecture links enhancers and non-coding disease variants to target gene promoters. *Cell* **167**, 1369–1384 (2016).
41. Aguet, F. et al. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
42. Han, H. et al. TRRUST v2: An expanded reference database of human and mouse transcriptional regulatory interactions. *Nucleic Acids Res.* **46**, D380–D386 (2018).
43. Thomsen, E.R. et al. Fixed single-cell transcriptomic characterization of human radial glial diversity. *Nat. Methods* **13**, 87-93 (2016).
44. Pollen, Alex A. et al. Molecular identity of human outer radial glia during cortical development. *Cell* **163**, 55-67 (2015).
45. Tran, H.T.N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 12 (2020).

46. Stark, S.G. et al. SCIM: Universal single-cell matching with unpaired feature sets. *Bioinformatics* **36**, i919–i927 (2020).
47. Yang, K.D. et al. Multi-domain translation between single-cell imaging and sequencing data using autoencoders. *Nat. Commun.* **12**, 31 (2021).
48. Bandura, D.R. et al. Mass cytometry: Technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Anal. Chem.* **81**, 6813–6822 (2009).
49. 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).
50. Ashuach, T., Reidenbach, D.A., Gayoso, A. & Yosef, N. PeakVI: A deep generative model for single cell chromatin accessibility analysis. Preprint at https://www.biorxiv.org/content/10.1101/2021.04.29.442020 (2021).
51. Hamilton, W., Ying, Z. & Leskovec, J. Inductive representation learning on large graphs. In Advances in Neural Information Processing Systems. (eds. I. Guyon et al.) 1024–1034 (Curran Associates, Inc., Long Beach, CA, USA, 2017).
52. Veličković, P. et al. Graph attention networks. Preprint at https://arxiv.org/abs/1710.10903 (2017).
53. Vashishth, S., Sanyal, S., Nitin, V. & Talukdar, P. Composition-based multi-relational graph convolutional networks. In *Proceedings of the 8th International Conference on Learning Representations*. (Addis Ababa, Ethiopia, 2020).
54. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat. Rev. Genet.* **20**, 257–272 (2019).
55. Amodio, M. & Krishnaswamy, S. MAGAN: Aligning biological manifolds. In *Proceedings of the 35th International Conference on Machine Learning*. (eds. J.G. Dy & A. Krause) 215–223 (PMLR, Stockholm, Sweden, 2018).
56. Tarashansky, A.J. et al. Mapping single-cell atlases throughout metazoa unravels cell type evolution. *eLife* **10**, e66747 (2021).
57. Ding, J. & Regev, A. Deep generative model embedding of single-cell RNA-seq profiles on hyperspheres and hyperbolic spaces. *Nat. Commun.* **12**, 2554 (2021).
58. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems. (eds. C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani & K.Q. Weinberger) 3111–3119 (Curran Associates, Inc., Lake Tahoe, NV, USA, 2013).
59. Kipf, T.N. & Welling, M. Semi-supervised classification with graph convolutional networks. In *Proceedings of the 5th International Conference on Learning Representations*. (Toulon, France, 2017).
60. Dincer, A.B., Janizek, J.D. & Lee, S.-I. Adversarial deconfounding autoencoder for learning robust gene expression embeddings. *Bioinformatics* **36**, i573–i582 (2020).
61. Goodfellow, I. et al. Generative adversarial nets. In Advances in Neural Information Processing Systems. (eds. Z. Ghahramani, M. Welling, C. Cortes, N.D. Lawrence & K.Q. Weinberger) 2672–2680 (Curran Associates, Inc., Montreal, Quebec, Canada, 2014).
62. Aibar, S. et al. SCENIC: Single-cell regulatory network inference and clustering. *Nat. Methods* **14**, 1083–1086 (2017).
63. Davis, C.A. et al. The encyclopedia of DNA elements (ENCODE): Data portal update. *Nucleic Acids Res.* **46**, D794–D801 (2018).

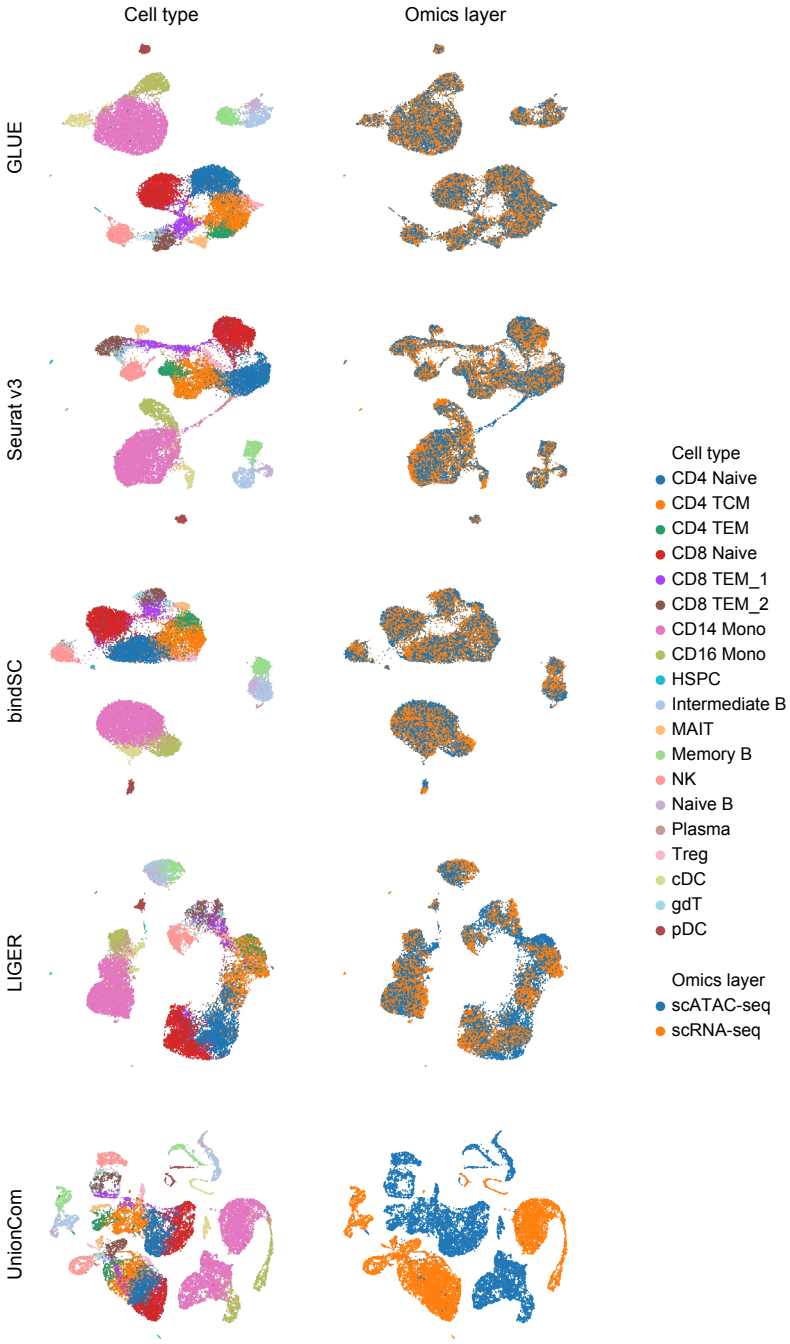
Supplementary Figures



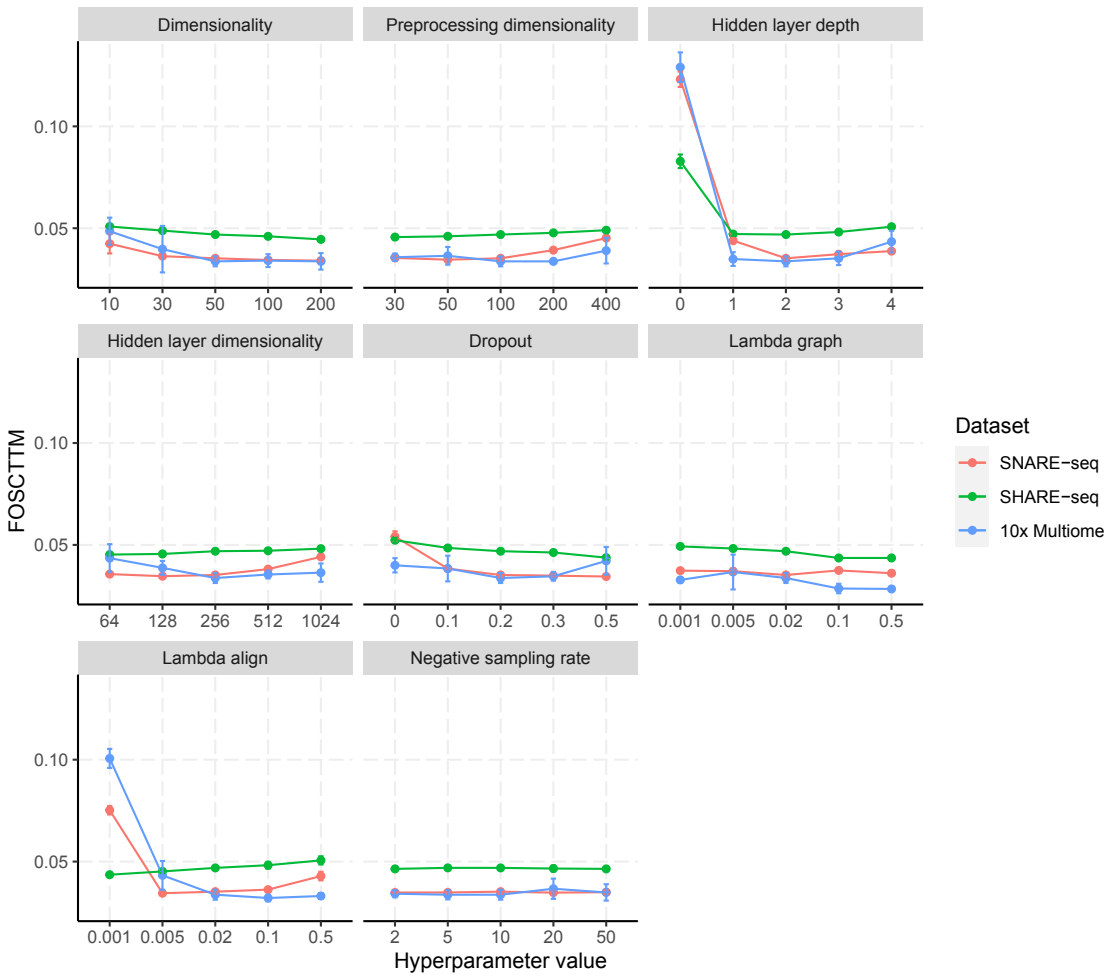
Supplementary Fig. 1 UMAP visualizations of the cell embeddings in the SNARE-seq dataset aligned with different integration methods.



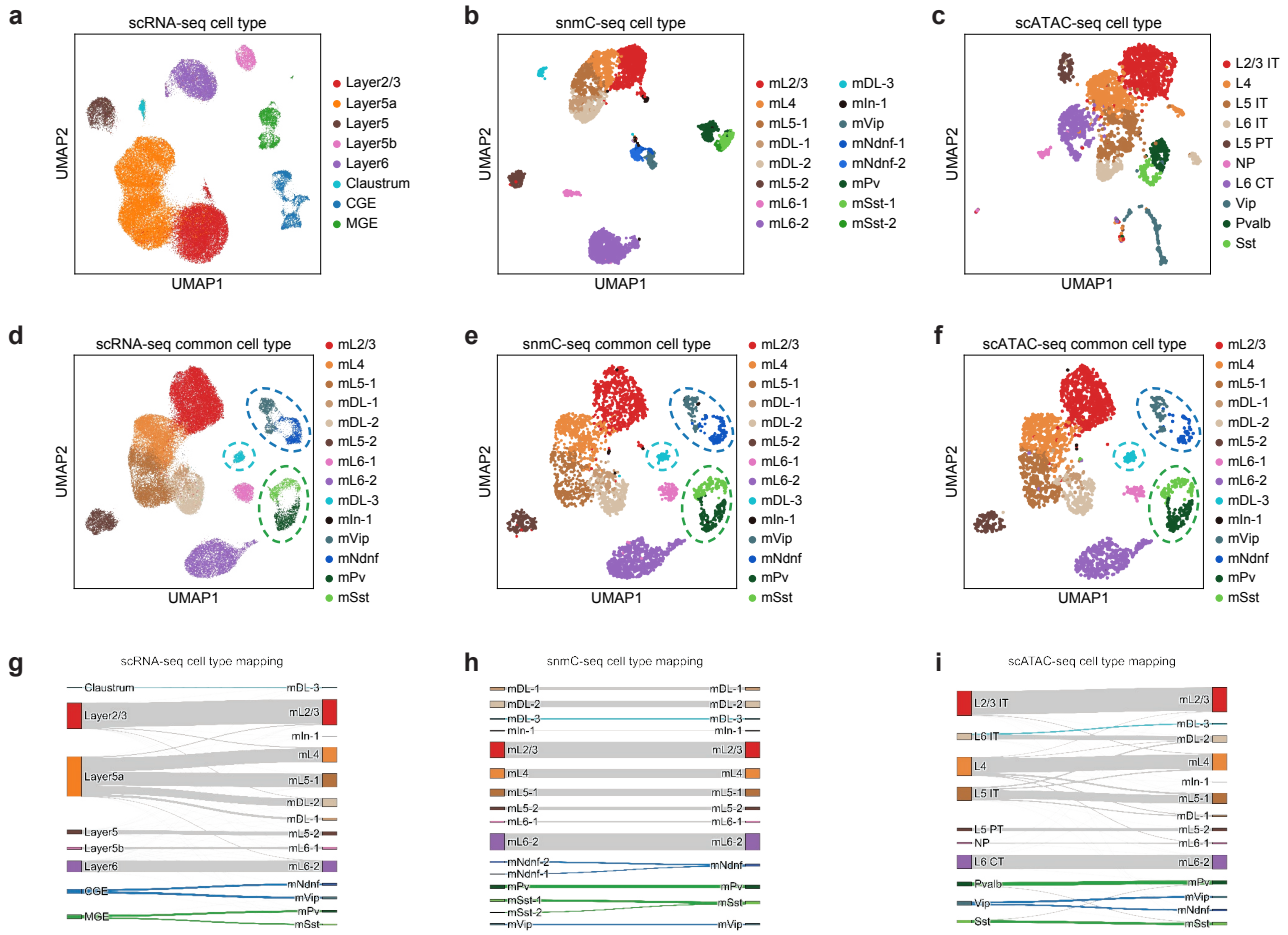
Supplementary Fig. 2 UMAP visualizations of the cell embeddings in the SHARE-seq dataset aligned with different integration methods.
UnionCom failed to run because of memory overflow.



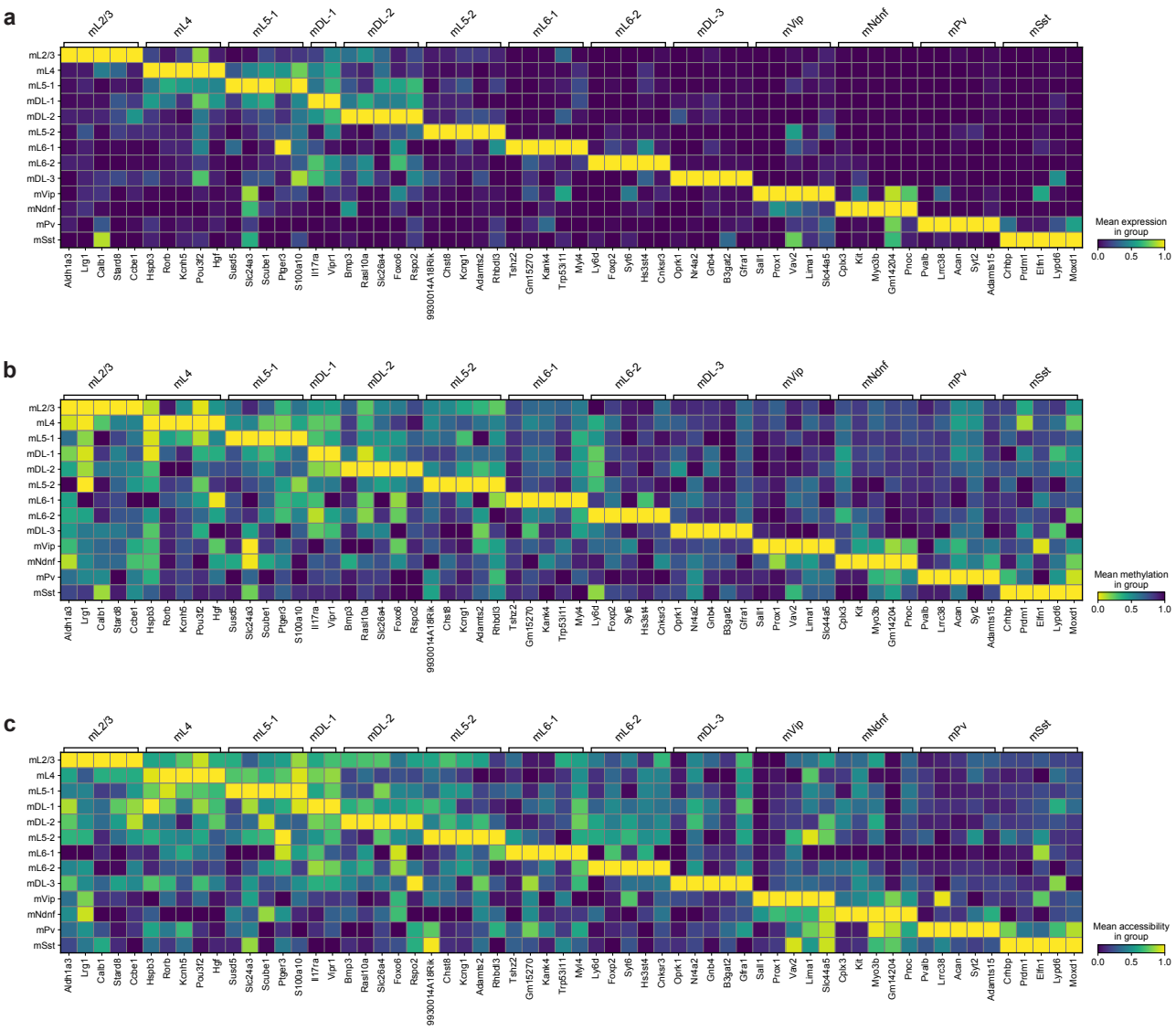
Supplementary Fig. 3 UMAP visualizations of the cell embeddings in the 10x Multiome dataset aligned with different integration methods.



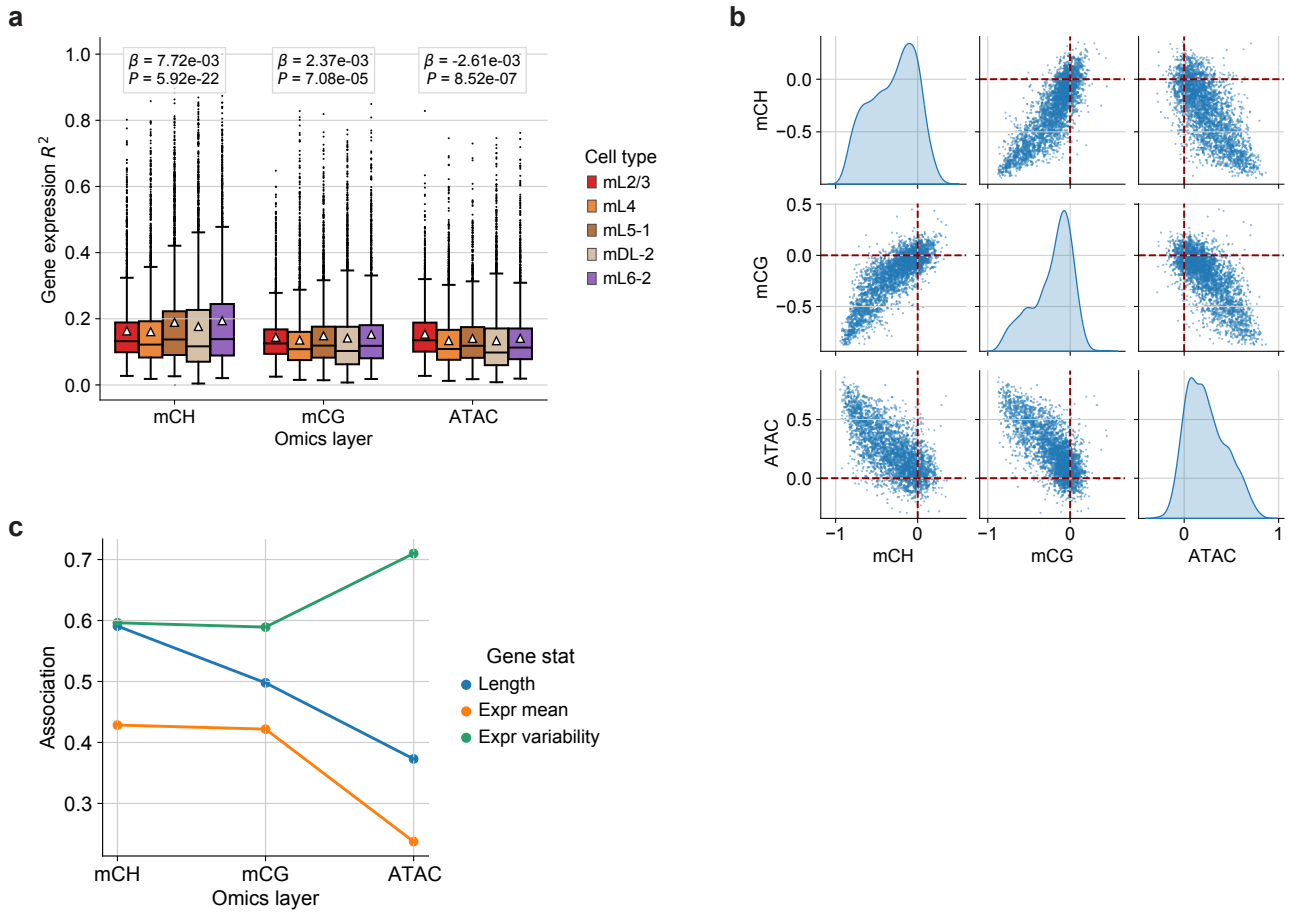
Supplementary Fig. 4 FOSCTTM of GLUE under different hyperparameter settings. "Dimensionality" denotes the cell embedding dimensionality. "Preprocessing dimensionality" is the reduced dimensionality used for the first transformation layers of the data encoders (see Methods). "Hidden layer depth" is the number of hidden layers in the data encoders and modality discriminator. "Hidden layer dimensionality" is the dimensionality of hidden layers in the data encoders and modality discriminator. "Dropout" is the dropout rate of hidden layers in data encoders and modality discriminator. "Lambda graph" is the weight of the graph loss ($\lambda_{\mathcal{G}}$). "Lambda align" is the weight of the adversarial alignment (λ_{D}). "Negative sampling rate" is the number of empirical samples used in negative edge sampling (samples from p_{ns}). For each hyperparameter, the center value is the default. To control computational cost, one hyperparameter was varied at a time, with all others set to their default values. The performance of GLUE was robust across a wide range of hyperparameter settings, except for failed alignments in which the adversarial alignment weight was too low or no hidden layers were used in the neural networks (equivalently a linear model with insufficient capacity).



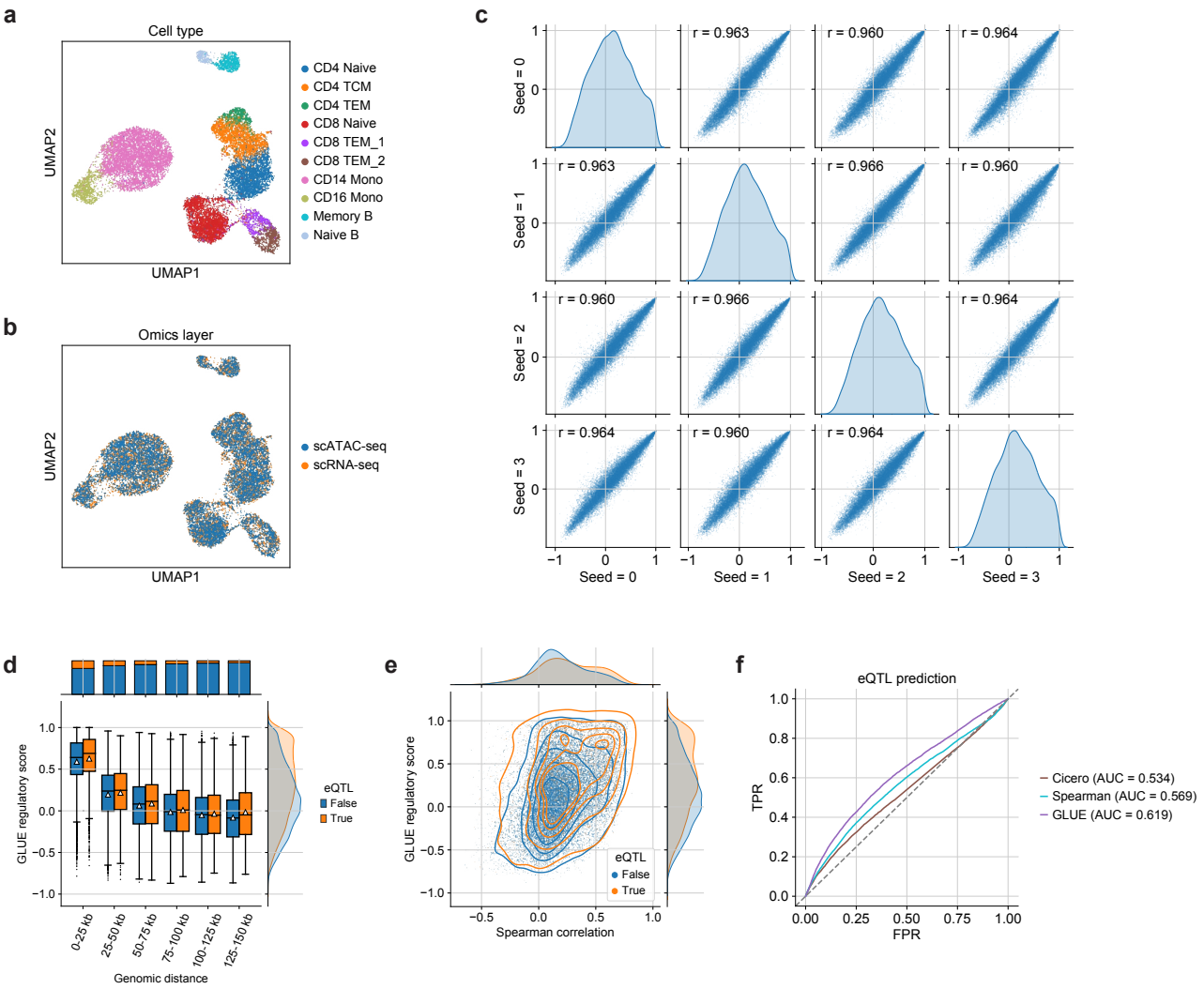
Supplementary Fig. 5 Triple-omics alignment and label transfer.
a-c, UMAP visualizations of the cell embeddings before integration for **a**, scRNA-seq, **b**, snmC-seq, and **c**, scATAC-seq, colored by the original cell types. **d-f**, UMAP visualizations of the integrated cell embeddings for **d**, scRNA-seq, **e**, snmC-seq, and **f**, scATAC-seq, colored by the unified cell types (labels transferred from snmC-seq). **g-i**, Alluvial diagrams comparing the original cell types and unified cell types for **g**, scRNA-seq, **h**, snmC-seq, and **i**, scATAC-seq. The original cell types are to the left, and the unified cell types are to the right. Cells relabeled to "mPv" and "mSst" are highlighted with green flows. Cells relabeled to "mNdnf" and "mVip" are highlighted with dark blue flows. Cells relabeled to "mDL-3" are highlighted with light blue flows.



Supplementary Fig. 6 Consensus cell type markers across three omics layers.
a, Expression in scRNA-seq. **b**, Gene body DNA methylation in snmC-seq. **c**, Chromatin accessibility in scATAC-seq. Note that an inverted color bar is used for DNA methylation.

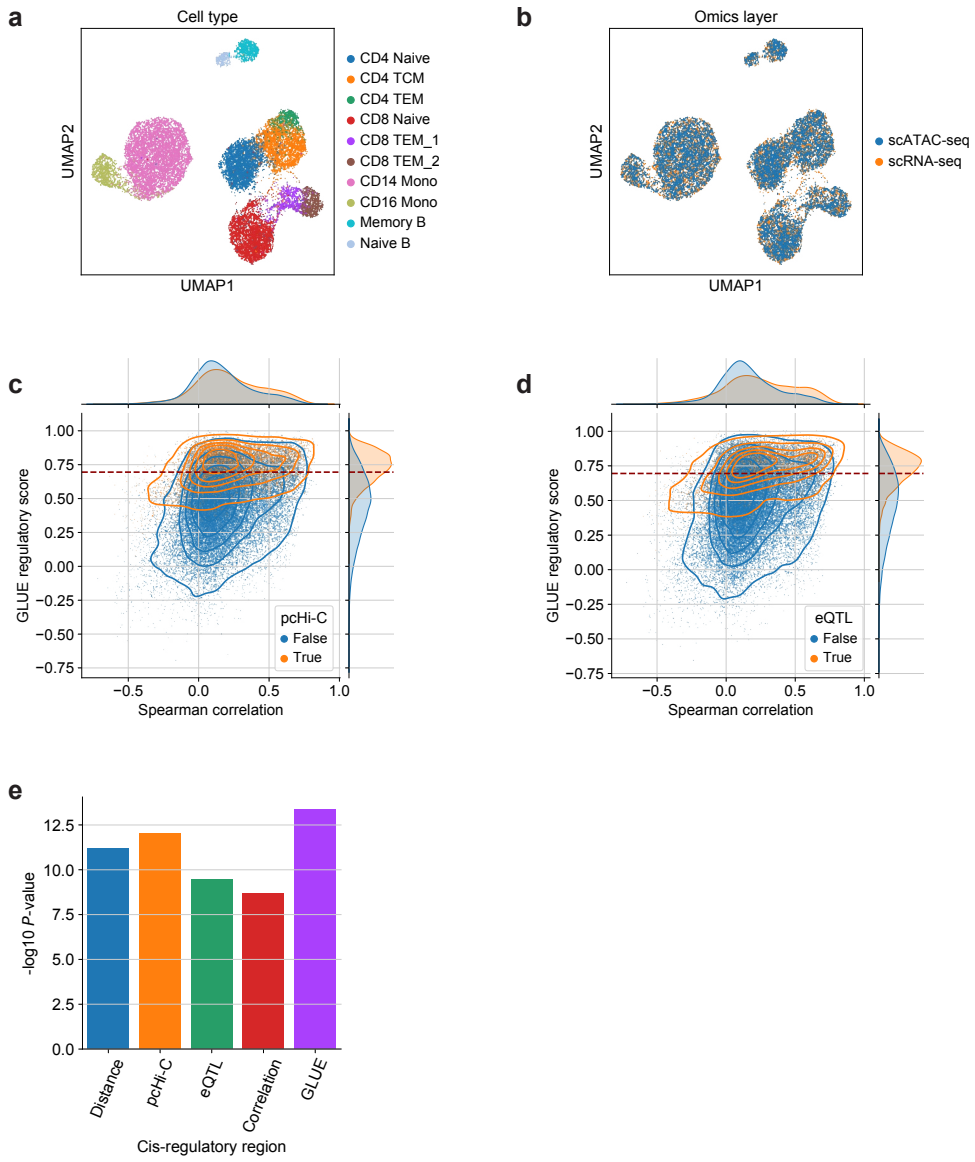


Supplementary Fig. 7 Epigenetic contributions to gene expression.
a, Coefficient of determination (R^2) for predicting gene expression based on each epigenetic layer in different cell types. The box plots indicate the medians (centerlines), means (triangles), 1st and 3rd quartiles (hinges), and minima and maxima (whiskers). Above each epigenetic layer, the linear regression slope (β) and its P value are displayed, with the cell type as the regressor. **b**, Pearson's correlations between different epigenetic states and gene expression. **c**, Associations (defined as the absolute values of Pearson's correlations) between the correlations in **b** and different gene characteristics.



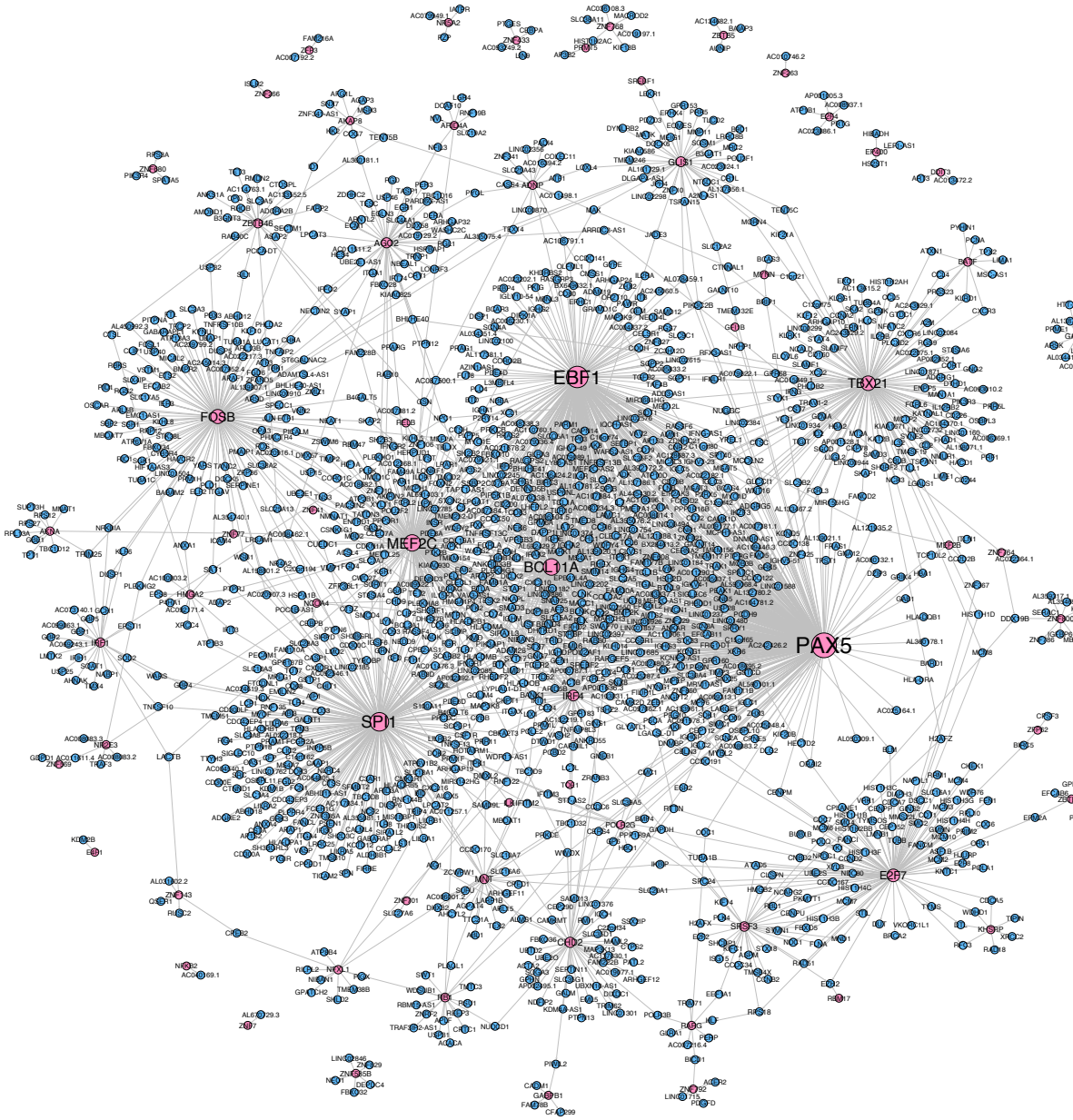
Supplementary Fig. 8 Model-based regulatory inference using distance-based power-law interactions as guidance.

a, **b**, UMAP visualizations of the integrated cell embeddings colored by **a**, cell types, and **b**, omics layers. **c**, Pearson's correlation coefficients for the GLUE regulatory scores across different random seeds. **d**, GLUE regulatory scores for peak-gene pairs across different genomic ranges, grouped by whether they had eQTL support. The box plots indicate the medians (centerlines), means (triangles), 1st and 3rd quartiles (hinges), and minima and maxima (whiskers). **e**, Comparison between the GLUE regulatory scores and the empirical peak-gene correlations computed on paired cells. Peak-gene pairs are colored by whether they had eQTL support. **f**, ROC (receiver operating characteristic) curves for predicting eQTL interactions based on different peak-gene association scores.

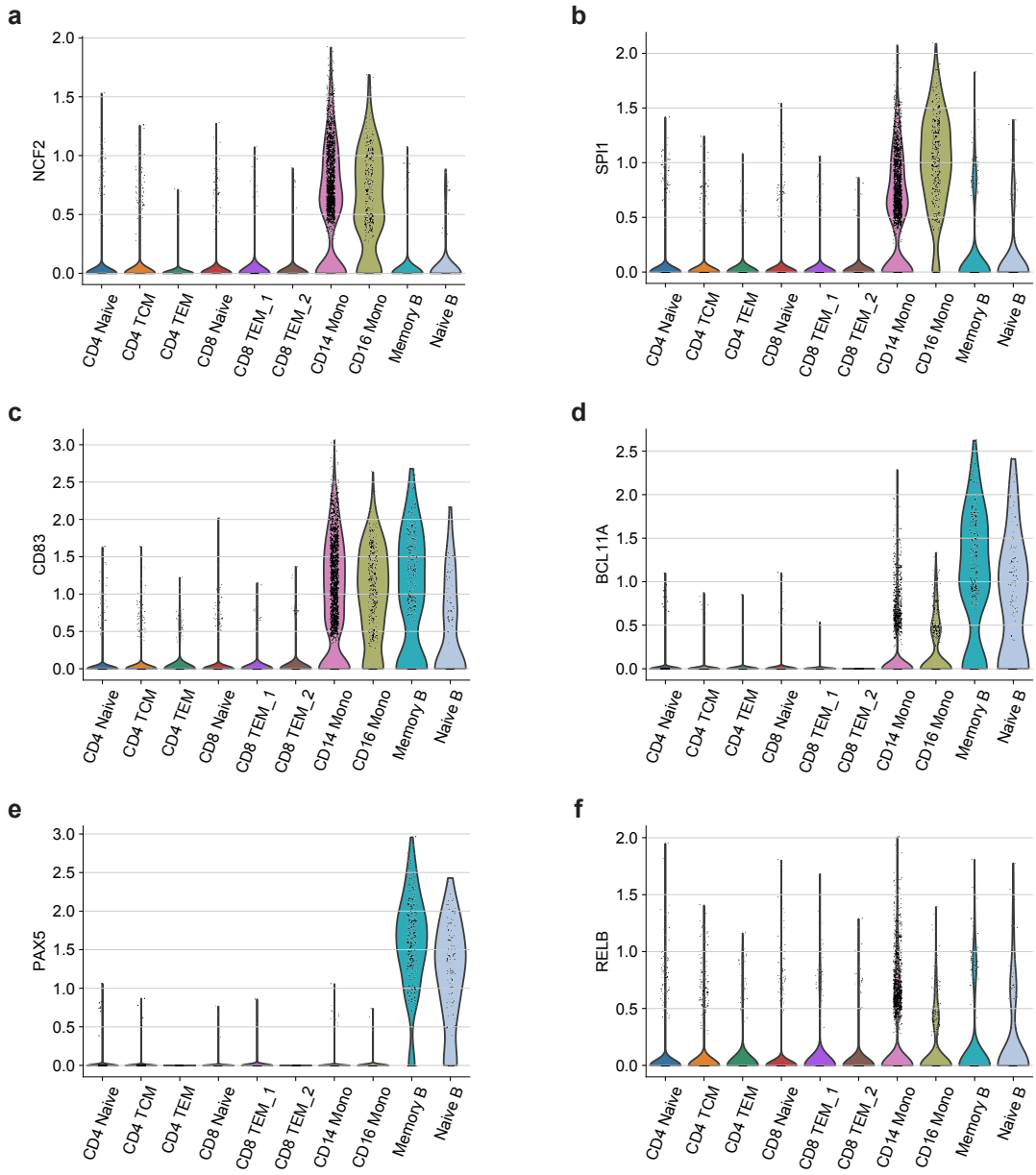


Supplementary Fig. 9 Model-based regulatory inference using a combination of distance-based, eQTL and pcHi-C interactions as guidance.

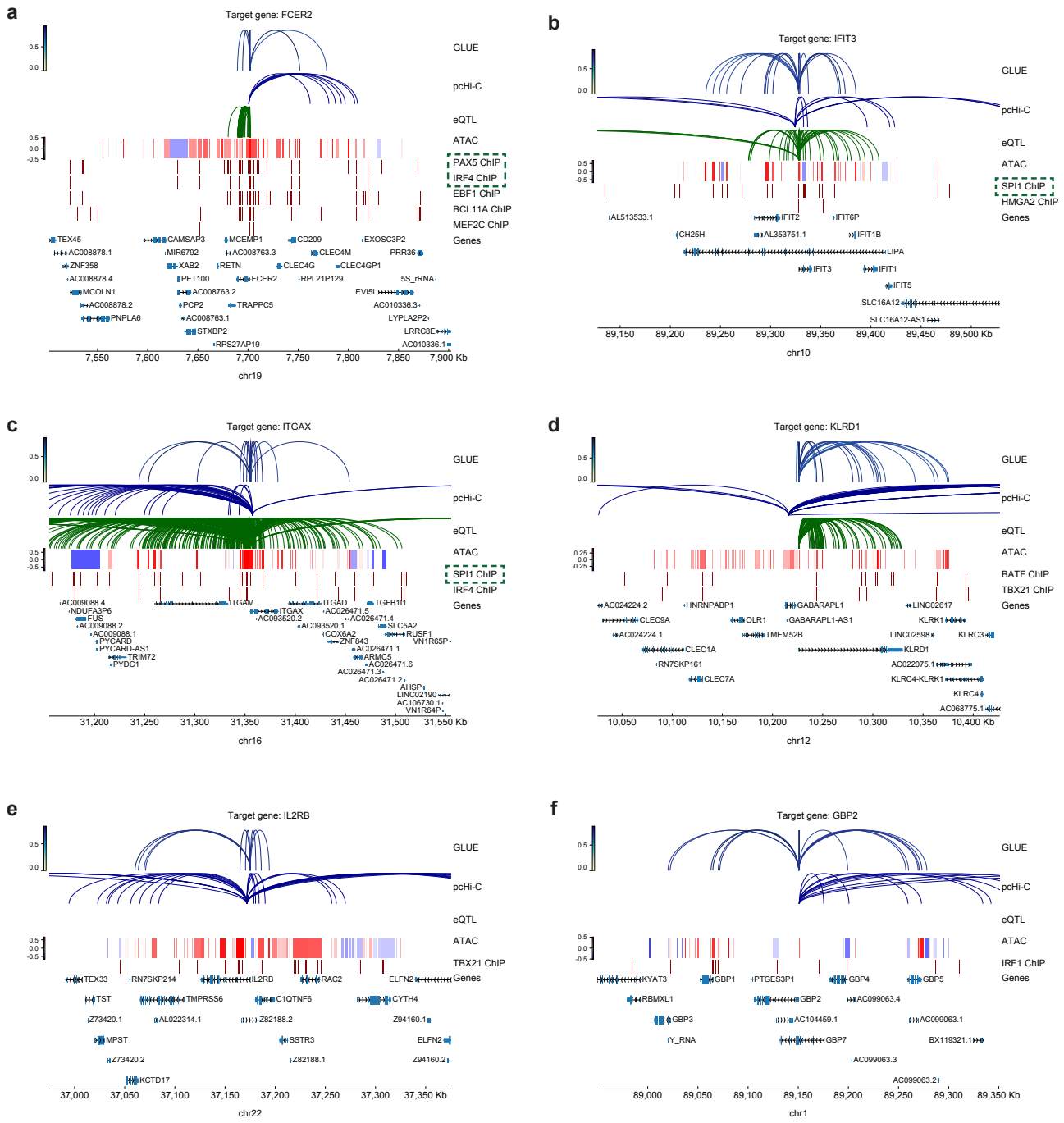
a, **b**, UMAP visualizations of the integrated cell embeddings colored by **a**, cell types, and **b**, omics layers. **c**, **d**, Comparison between the GLUE regulatory scores and the empirical peak-gene correlations computed on paired cells. The peak-gene pairs are colored by whether they had **c**, pcHi-C support or **d**, eQTL support. Red dashed lines indicate the 75th percentile of the GLUE regulatory scores, which was used as a cutoff. The GLUE-identified interactions mostly exhibited positive empirical correlations and covered the majority of pcHi-C and eQTL support, while the selection of peak-gene pairs based solely on empirical correlation would lead to much lower external support. **e**, Consistency between the TF-target gene networks constructed with different peak-gene association methods and the manually curated connections in the TRRUST v2 database (Fisher's exact test).



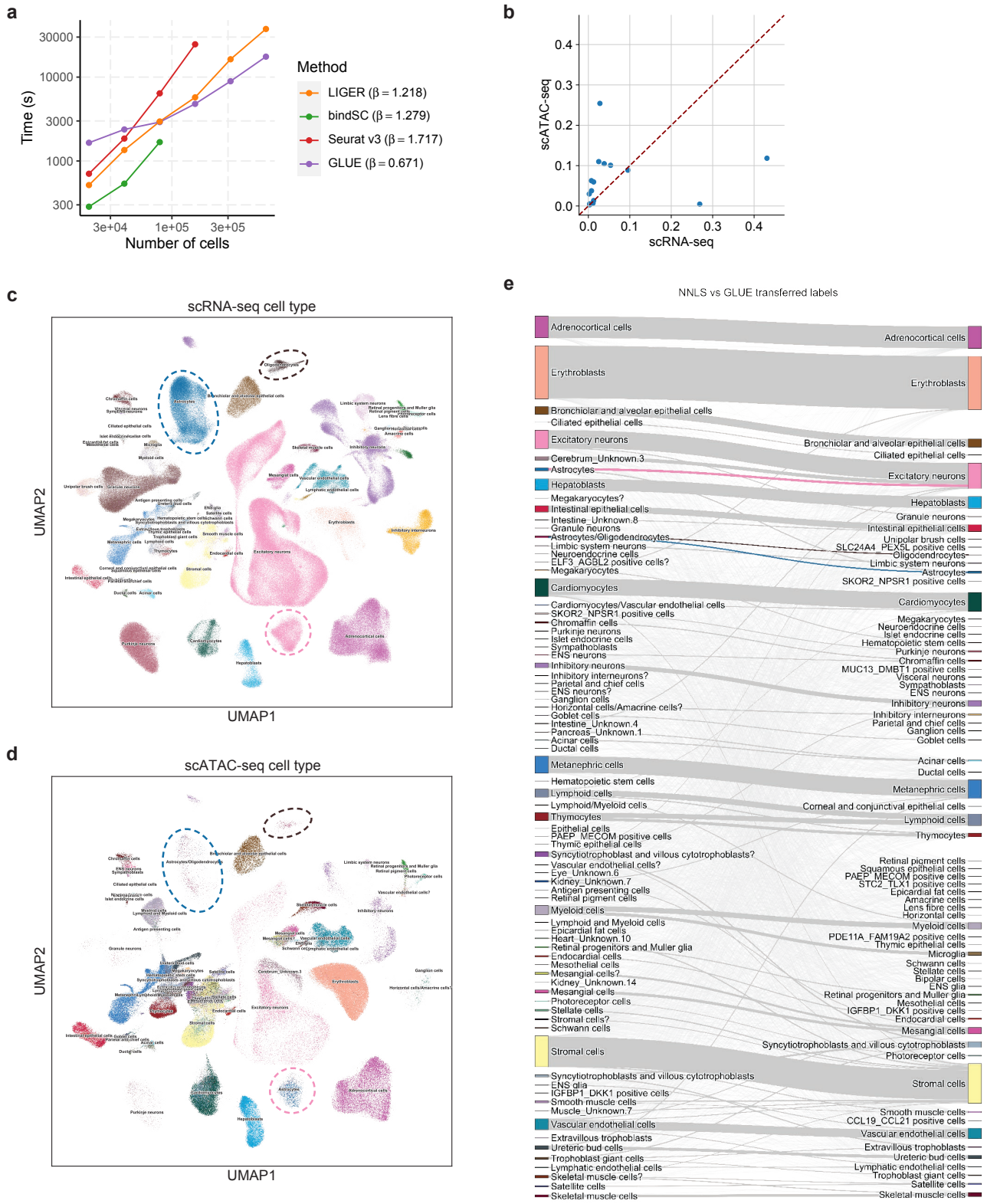
Supplementary Fig. 10 Inferred TF-target gene regulatory network in PBMC.



Supplementary Fig. 11 TF-target gene expression levels in different cell types.
a, **b**, Expression levels of **a**, *NCF2*, and its regulator **b**, *SPI1* in different cell types. **c**-**f**, Expression levels of **c**, *CD83*, and its inferred regulators **d**, *BCL11A*, **e**, *PAX5*, and **f**, *RELB* in different cell types.



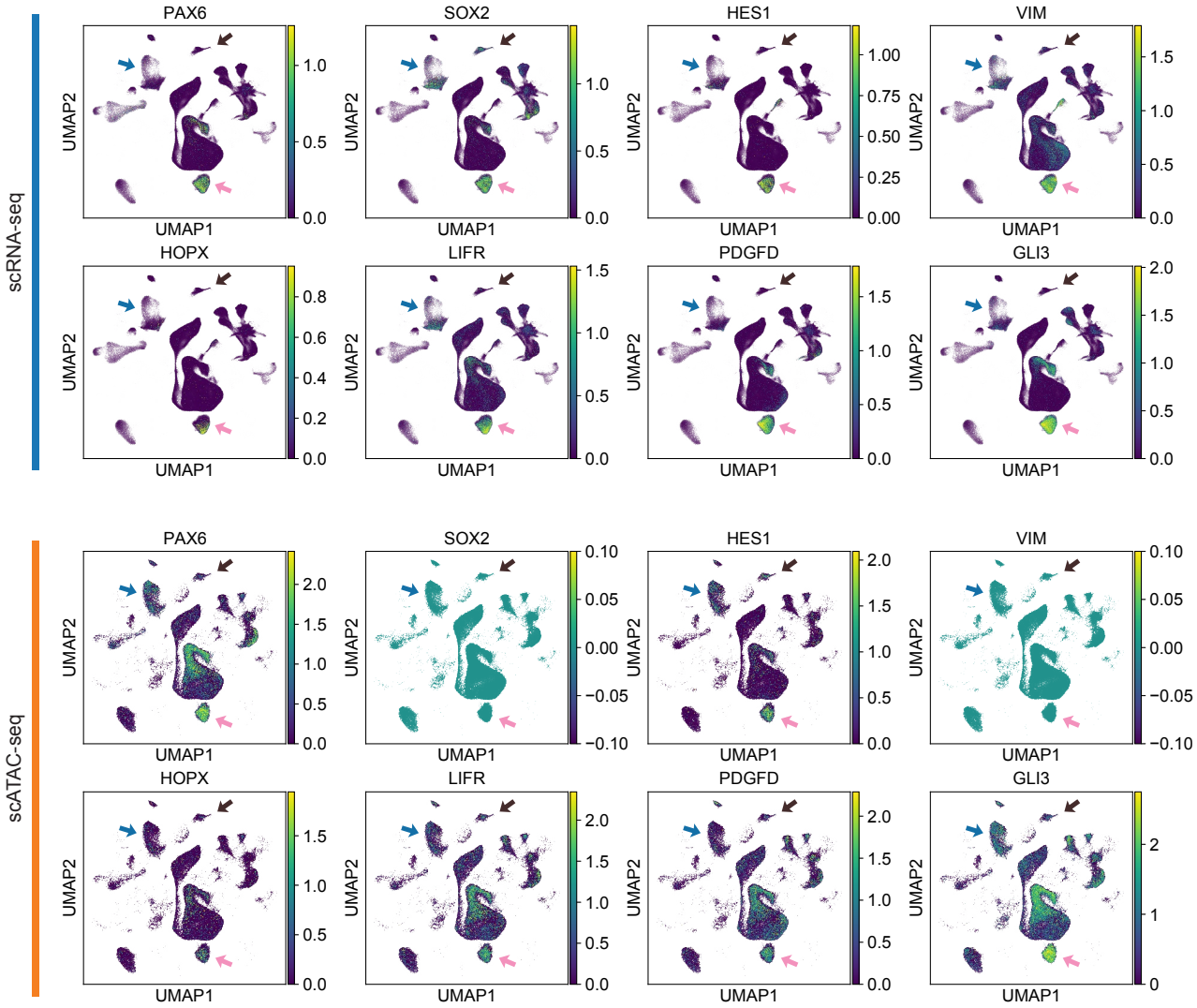
Supplementary Fig. 12 Regulatory inference and evidence per gene.
GLUE-identified cis-regulatory interactions for **a**, *FCER2*, **b**, *IFIT3*, **c**, *ITGAX*, **d**, *KLRD1*, **e**, *IL2RB*, and **f**, *GBP2*, along with individual pieces of regulatory evidence. For each gene, the ChIP-seq tracks correspond to inferred TF regulators. Known regulators are highlighted with green boxes.



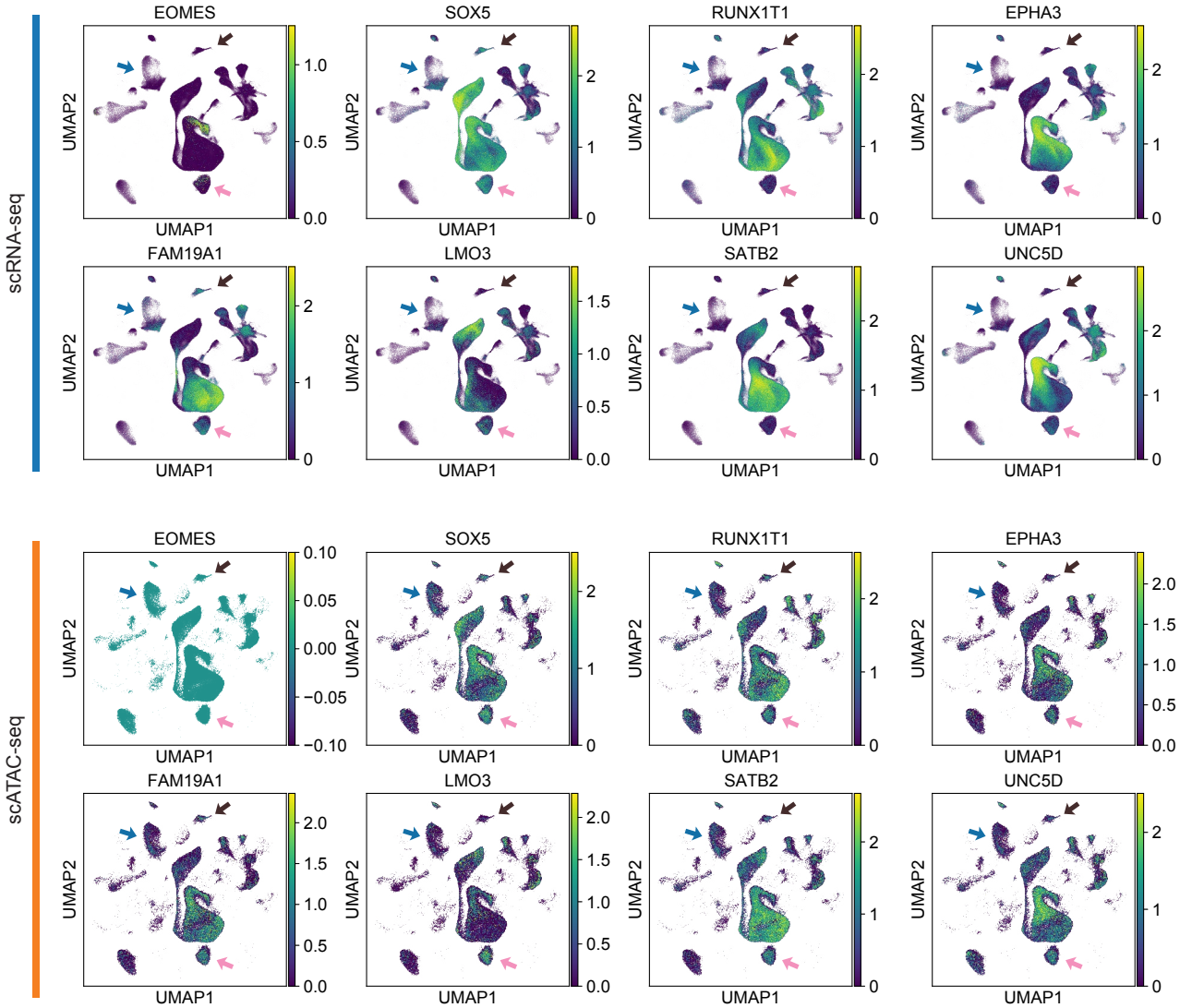
(Legend on next page)

Supplementary Fig. 13 Scalability benchmarking and integration of a multi-omics human cell atlas.

a, Time costs of different methods on subsampled atlases of varying sizes. The time costs and cell numbers are both plotted in log-scale, so the slope β represents the exponent in original scale, i.e., $\beta = 2$ represents quadratic scalability, $\beta = 1$ represents linear scalability, and $\beta < 1$ represents sublinear scalability. **b**, Organ compositions in scRNA-seq and scATAC-seq. **c**, UMAP visualizations of the integrated cell embeddings showing only the scRNA-seq cells, colored by the original cell types. **d**, UMAP visualizations of the integrated cell embeddings showing only the scATAC-seq cells, colored by the original cell types. **e**, Alluvial diagram comparing the NNLS (non-negative least squares)-based and GLUE-based cell type annotations. The original NNLS-based annotations are to the left, and the GLUE-based annotations are to the right. Cells originally labeled as "Astrocytes" but mapped to "Excitatory neurons" are highlighted with pink flows. Cells originally labeled as "Astrocytes/Oligodendrocytes" but mapped to "Astrocytes" are highlighted with blue flows. Cells originally labeled as "Astrocytes/Oligodendrocytes" but mapped to "Oligodendrocytes" are highlighted with brown flows.

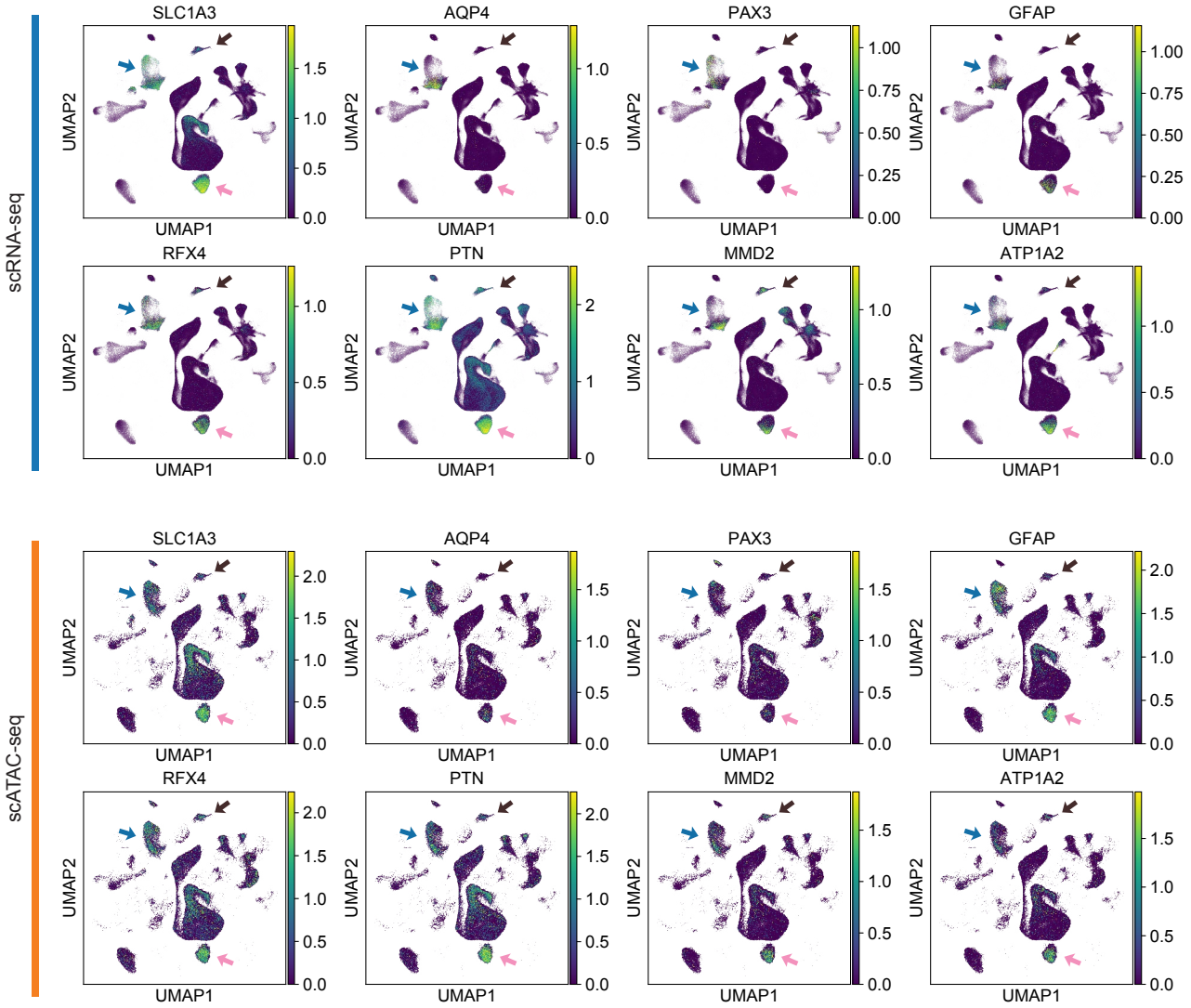


Supplementary Fig. 14 Gene expression and chromatin accessibility patterns of neural progenitor markers in cerebrum cells.
Putative neural progenitors are highlighted with pink arrows, astrocytes are highlighted with blue arrows, and oligodendrocytes are highlighted with brown arrows. *SOX2* and *VIM* were not detected in scATAC-seq, due to limitations in ATAC gene-level score calculation (no peaks overlap gene body and 2 kb upstream from TSS).

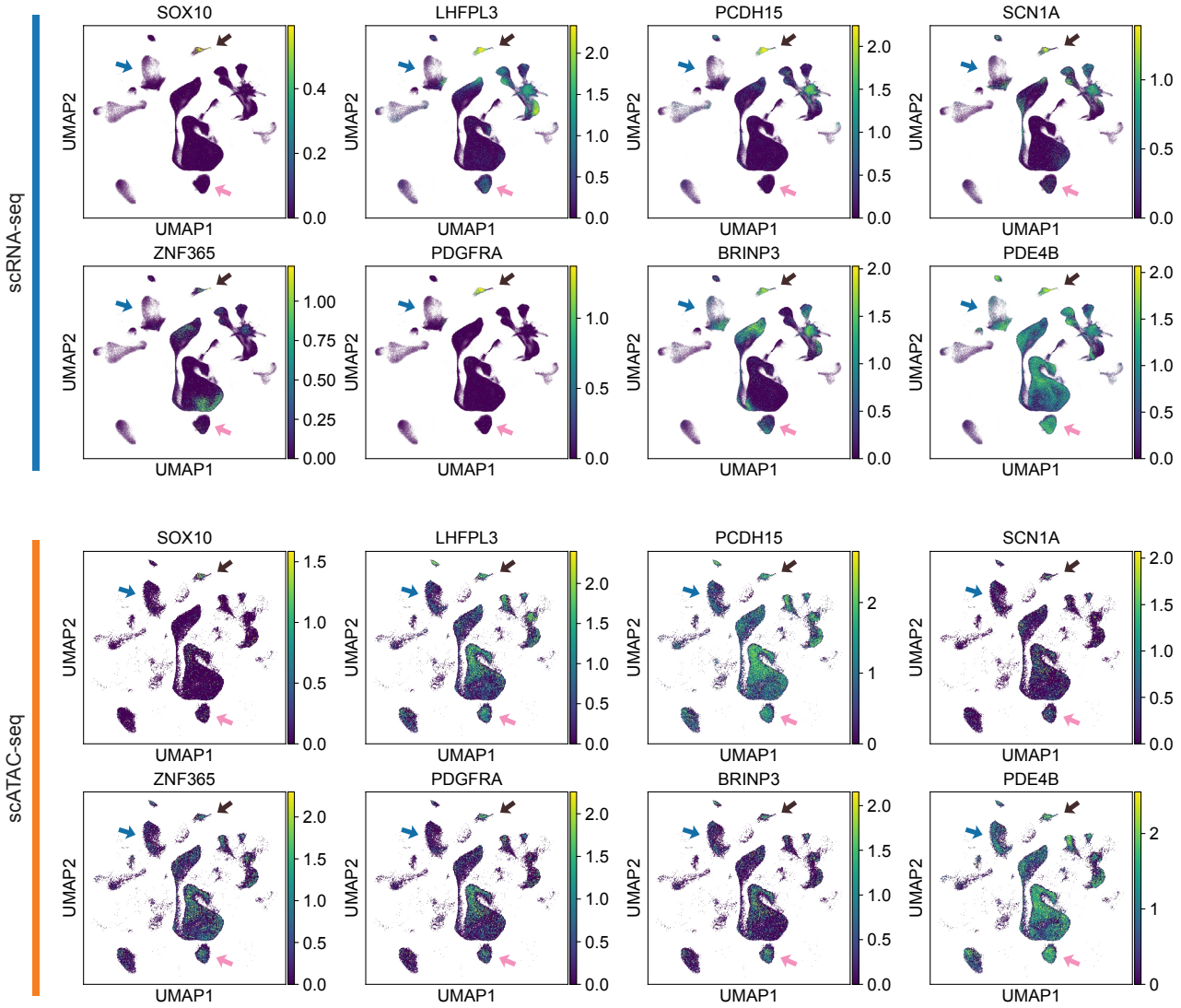


Supplementary Fig. 15 Gene expression and chromatin accessibility patterns of excitatory neuron markers in cerebrum cells.

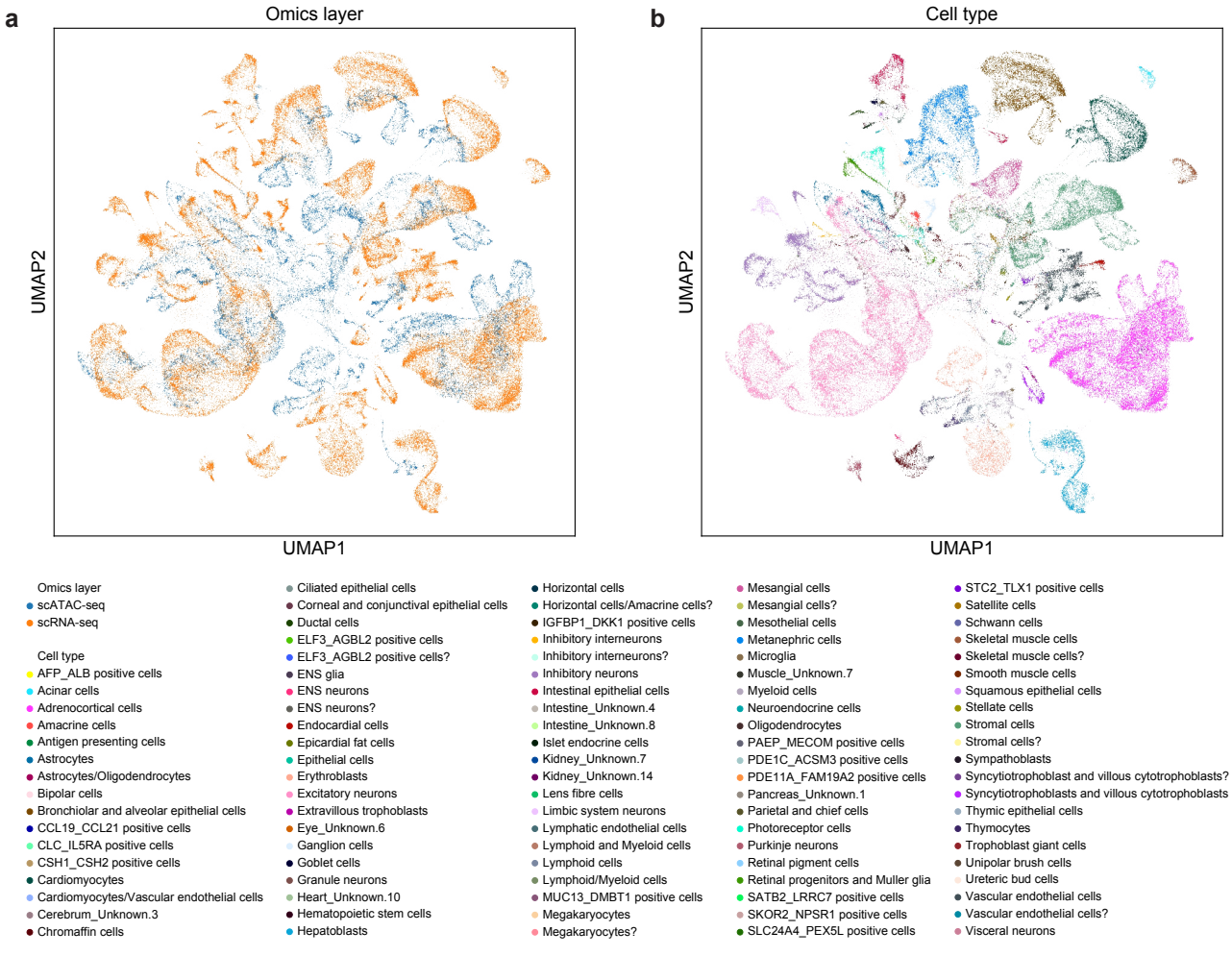
Putative neural progenitors are highlighted with pink arrows, astrocytes are highlighted with blue arrows, and oligodendrocytes are highlighted with brown arrows. *EOMES* was not detected in scATAC-seq, due to limitations in ATAC gene-level score calculation (no peaks overlap gene body and 2 kb upstream from TSS).



Supplementary Fig. 16 Gene expression and chromatin accessibility patterns of astrocyte markers in cerebrum cells.
Putative neural progenitors are highlighted with pink arrows, astrocytes are highlighted with blue arrows, and oligodendrocytes are highlighted with brown arrows.



Supplementary Fig. 17 Gene expression and chromatin accessibility patterns of oligodendrocyte markers in cerebrum cells.
Putative neural progenitors are highlighted with pink arrows, astrocytes are highlighted with blue arrows, and oligodendrocytes are highlighted with brown arrows.



Supplementary Fig. 18 UMAP visualization of the Seurat v3 integration of the aggregated multi-omics human cell atlas.
Coembeddings of the aggregated metacells colored by **a**, omics layers, and **b**, cell types.

Supplementary Table Legends

Supplementary Table 1 Public datasets used in the study.

Supplementary Table 2 Detailed benchmark data.

Supplementary Table 3 Regulatory interactions in the GLUE-derived TF-target gene network.