STEP: Spatial Transcriptomics Embedding Procedure for Multi-scale Biological Heterogeneities Revelation in Multiple Samples

Lounan Li [1,2,†], Zhong Li [1,2,†,*], Yuanyuan Li[3], Xiao-ming Yin[3], Xiaojiang Xu [3,*]

1. School of Information Engineering, Huzhou University, Huzhou, Zhejiang, 313000, China

2. College of Science, Zhejiang Sci-Tech University, Hangzhou, Zhejiang, 310018, China

3. School of Medicine, Tulane University, New Orleans, LA, 70112, USA

† Co-first author. * Co-corresponding author. Contact: xxu17@tulane.edu

Abstract

In the realm of spatially resolved transcriptomics (SRT) and single-cell RNA sequencing (scRNA-seq), addressing the intricacies of complex tissues, integration across non-contiguous sections, and scalability to diverse data resolutions remain paramount challenges. We introduce STEP (Spatial Transcriptomics Embedding Procedure), a novel foundation AI architecture for SRT data, elucidating the nuanced correspondence between biological heterogeneity and data characteristics. STEP's innovation lies in its modular architecture, combining a Transformer and β-VAE based backbone model for capturing transcriptional variations, a novel batch-effect model for correcting inter-sample variations, and a graph convolutional network (GCN)-based spatial model for incorporating spatial context—all tailored to reveal biological heterogeneities with unprecedented fidelity. Notably, STEP effectively scales to the newly proposed 10x Visium HD technology for both cell type and spatial domain identifications. STEP also significantly improves the demarcation of liver zones, outstripping existing methodologies in accuracy and biological relevance. Validated against leading benchmark datasets, STEP redefines computational strategies in SRT and scRNA-seq analysis, presenting a scalable and versatile framework for dissecting complex biological systems.

Introduction

Developments in spatially resolved transcriptomics (SRT) and single-cell RNA sequencing (scRNA-seq) have revolutionized our comprehension of gene expression patterns. SRT, a fusion of histology and transcriptomics, offers precise transcript mapping within tissues, unveiling unprecedented details about tissue structure and gene expression profiles. Technologies like Visium (Stahl et al., 2016) and Slide-seq V2 (Stickels et al., 2021) have facilitated the collection of data at a non-single-cell resolution, capturing gene expression in 'spots' comprising multiple cells. Analyzing such data presents significant challenges in distinguishing various biological heterogeneities within these multicellular mixtures. In contrast, methodologies such as Visium HD (Nagendran et al., 2023), MERFISH (Chen et al., 2015; Moffitt et al., 2016) and STARmap (Wang et al., 2018) provide single-cell resolution in spatial transcriptomics,

offering cell-specific gene expression data. These approaches play a crucial role in uncovering individual cellular identities and interactions, thereby enhancing our understanding of complex biological systems.

Accordingly, various computational tools are developed to solve these categories of problems in SRT data analyses: 1. Cell clustering, aiming to differentiate cells by cell types (Chidester et al., 2023; Li et al., 2022; Miller et al., 2021; Teng et al., 2022); 2. Spatial domain identification for discovery of biologically functional regions with a certain degree of spatial continuity in tissues (Dong & Zhang, 2022; Hu et al., 2021; Hu et al., 2024; Li & Zhou, 2022; Liu et al., 2023; Long et al., 2023; Shang & Zhou, 2022; Yang et al., 2024; Yu et al., 2023; Zhang et al., 2024; Zhao et al., 2021). The significant improvements can be observed on solving these two tasks in the single section scenario for a specific data resolution. Conversely, these tools are always unable to adapt to the diverse data resolutions produced by different SRT technologies. More importantly, the intricacies of the complex tissues whose regions are always unstructured and in a low level of spatial continuity, present a substantial challenge for extant analytical methods. This can be particularly highlighted by liver zonation. Addressing this biological conundrum necessitates a methodology adept at not just capturing the full complexity of transcript information but also proficiently harnessing spatial context. Furthermore, understanding the complex mechanisms and characterizing the tissue environment require integrating data from multiple sections. These sections come from different patients and cover various regions and conditions, such as distinctive treatments for the same ailment and genetic modifications like gene knockouts. This reconciliation calls for computational methods capable of integrating spatial transcriptomic data from such disparate and non-contiguous slices. Regrettably, the current suite of tools often falls short in meeting these multifaceted demands.

These limitations of current methods arise from their excessive focus on specific tasks or data characteristics—namely, whether the data have single-cell resolution ("resolution orientation") or are spatially resolved ("modality orientation"). These tendencies ignore the correspondence between the target underlying biological heterogeneities and source data resolutions and modalities. This vagueness further leads to the excessive focus on spatial variations but the underestimation of essential difference at transcriptional level. For example, some methods emphasize the correlation between spatial distances and similarity of cell types as the core assumption of their cell clustering methods (Li et al., 2022; Miller et al., 2021). Not only are these assumptions incompletely correct, they are actually unnecessary because cell types can be well distinguished at the level of gene expression regardless of spatial locations. On the other hand, when it comes to the integration of multiple sections, current methods often require the spatial contiguity between sections, whether vertically or horizontally (Clifton et al., 2023; Dong & Zhang, 2022; Gao et al., 2024; Long et al., 2023; Wang et al., 2023; Yu et al., 2023; Yuan, 2024; Zhou et al., 2023).Though, in concept, spatial domains demonstrate the spatial variations and the lower resolution compared to the cell types, the most fundamental variations that segment the regions are still transcriptional. Although one Bayesian model (Li & Zhou, 2022) and a cell-type annotation based method (Yuan,

2024) can handle both the multi-scale or multi-section (non-contiguous) SRT dataset, there is another limitation unique to these methods: they are unable to provide embeddings that are consistent to their identified spatial domains for the downstream analyses, like pseudotime analysis, where the PCA and other batch-effect removal algorithms specific for scRNA-seq data (Haghverdi et al., 2018; Korsunsky et al., 2019) are conducted, instead. Such compromise actually leaves the spatial context aside the further biological explanations of the results. All in all, these causes can be concluded as the inability of existing methods to balance the transcriptional, batch and spatial variations in different scales of biological heterogeneities. And to resolve this imbalance issue, a universal architecture is required to model the SRT data.

Based on the above discussions, the design of our solution commences by the clarification of the correspondences between target biological heterogeneities, source data resolution and source data modalities. Specifically, the scales of biological heterogeneities can be categorized to the cell type level and spatial domain level, and these two heterogeneities can be bound to data resolution at single-cell level and data modality of the spatially resolved, respectively. Moreover, the relationships between biological heterogeneities and variations can be also well-established: with the transcriptional variation being the most essential one, the revelation of the cell type level heterogeneity requires the conservation of transcriptional variations but the elimination of batch variations; while of the spatial domain level one requires an extra incorporation of spatial variations.

With these foundational insights, we introduce STEP (Spatial Transcriptomics Embedding Procedure), a pioneering foundation deep learning architecture for SRT data that utilizes a "decoupled but combinable" (modular) strategy to effectively tackle the mentioned limitations. Our solution integrates three specialized models: 1. A Transformer (Vaswani et al., 2023) and β-VAE based backbone model (BBM), the core of STEP, designed for the precise extraction of transcriptional variations into low-dimensional embeddings, ensuring the scalability across diverse data resolutions; 2. A batch-embedding based paired transformation model (BEM), introducing an unparalleled batch-correction technique that operates independently from BBM and delineates a new direction in handling batch variations; and 3. A spatial model (SpM) leveraging the graph convolutional network (GCN) (Kipf & Welling, 2017) distinctively applied to low-dimensional embeddings as a spatial filter or smoother rather than direct gene expression data as a feature extractor (Dong & Zhang, 2022; Gao et al., 2024; Hu et al., 2021; Long et al., 2023; Yu et al., 2023; Zhang et al., 2024; Zhou et al., 2023)—a first in the field. This innovative use of GCN significantly reduces noise, effectively capturing spatial contexts and variations with enhanced biological relevance. Because, this usage releases the power of GCN for its naturality as a spatial filter or smoother (NT et al., 2019), allowing the major attentions still on the transcriptional features rather being biased seriously by spatial variations. Our decoupled strategy ensures a focused extraction of transcriptional variations, achieving a balanced attentions across transcriptional, batch, and spatial variations. On the other side, the STEP's combinability allows for tailored analysis across diverse data resolutions and modalities according to the target

heterogeneities to reveal, thus delineating multi-scale biological heterogeneities into consistent and comparative embeddings. Based on this unprecedented consistency and comparability, we also design an embedding-based cell type deconvolution algorithm with a spatial-domain-wise regularization, targeting cell type level heterogeneity in non-single-cell resolution SRT data.

To demonstrate STEP's versatility, we evaluate its performance across five distinct tasks with the combinations of three models based on target heterogeneities: 1. Integration of multiple batches to identify unified cell types or states. STEP is applied to both scRNA-seq data and single-cell resolution SRT data from multiple batches or platforms in this task, and adeptly handles clustering and batch effect removal in both embeddings and gene expressions. STEP achieves superior performance by employing the combo of BBM and BEM, outperforming existing methods. 2. Analysis of single-batch non-single-cell resolution SRT data. In this scenario, STEP, with combo of BBM and SpM, not only accurately identifies zones or regions within a tissue, but also exhibits highly enhanced spatial continuity, smoothness and developmental trajectory to current methodologies. 3. Integration of multi-batch non-single-cell resolution SRT data. For this kind of SRT dataset, by using the combo of all three models, STEP efficiently aligns both contiguous and non-contiguous sections into a shared latent space, and also outlines developmental trajectories for unified domains. This alignment enables comparative studies across tissues, facilitating the identification and exploration of consistent biological features among different spatial domains under different conditions. 4. Integration of multi-batch single-cell resolution SRT data. For single-cell resolution SRT data, STEP can not only identify cell-types but can also aggregate the cell embeddings to identify the spatial domains based on the given spatial context, revealing multi-scale consistent biological heterogeneities. 5. By merging scRNA-seq and SRT data, STEP achieves proficient co-embedding and cell-type deconvolution. This approach amalgamates the strengths of both data modalities, enhancing our understanding of complex biological structures. STEP is implemented based on the workflow SCANPY (Wolf et al., 2018) in an extendable way, and is open source with comprehensive tutorials available. Considering the architecture design of STEP, it has a great potential to scale to a Large Transcriptomic Model (LTM).

Results

STEP's architecture and workflows

The STEP's backbone model (BBM) is implemented based on Transformer and conceptualized on two fundamental biological principles: 1. Genes seldom operate in isolation but instead function as part of a group or module, such as a pathway. These gene modules comprise sets of genes collaborating in a coordinated manner to execute specific biological functions (Subramanian et al., 2005; Wang et al., 2008). 2. Their behavior is intricately regulated through interactions with other genes within the same module and also with other gene modules. These regulatory interactions among gene

modules involve diverse and complex mechanisms (Isaacs et al., 2003; Lemoine et al., 2021; Saelens et al., 2018; Segal et al., 2003). BBM extends these principals by two corresponding key steps. First, it maps gene expression data into a gene module sequence, referring to multi-view learning methods (Fan et al., 2021; Yu et al., 2020). Second, it captures the interactions of gene modules with self-attention mechanisms, the core of Transformer. In this way, BBM ensures itself the capability to extract complex transcriptional variations in a regulatory view. During its forward propagation process, BBM maps each z-scored gene expression data (a cell/spot) to a sequence of gene modules through learnable linear transformations, then processes gene module sequences with several Transformer encoders. We extract the outputs from BBM's bottleneck layer as the low-dimensional embedding of each gene expression dataset. To further process these representations, we utilize an architecture resembling β-VAE (Burgess et al., 2018) and incorporating the Zero Inflated Negative Binomial (ZINB) distribution to fit the raw gene expression data where ZINB distribution has been corroborated to be highly effective in modeling gene-expression data. BBM subsequently performs specialized standardization, i.e., reparameterization trick, on the low-dimensional representations by fitting them in standard normal distribution and reconstructs them in the original data space via the decoder (Fig. 1a).

Our newly proposed paired batch-effect model (BEM), drawing insights from current Pearson residual based normalization methods (Hafemeister & Satija, 2019; Lause et al., 2021) and the layer normalization (Lei Ba et al., 2016) in deep learning , offers a new way to handle batch effects. By pairing such normalization with an "inverted" operation before the fitting of raw gene-expression data, BEM comprises two core components: 1. Batch Encoder with "BatchAwareLayerNorm": the initial part addresses batch variations by modeling batch-dependent shifts and scales on observed gene expression data; 2. Batch Decoder with "BatchAwareScale": analogously, an element-wise scale transformation in the latent space is applied before the decoding process (Fig.1a). This BEM independently extends the BBM with a learnable batch-embedding concept, akin to essential word-embedding techniques widely used in natural language processing (NLP) methods. With this batch-embedding concept, the batch covariate can be effectively incorporated or regressed out for different purposes in a way akin to conditional generation (Supplementary Notes 1).

For the Batch Encoder, two MLPs use batch-embedding to estimate the batch-dependent shifts and scales of cells in each batch, performing element-wise addition and multiplication on the cell-wise z-scored gene expression data. This process is termed as "BatchAwareLayerNorm"; The Batch Decoder performs a batch-dependent element-wise scaling in the latent space, referred as "BatchAwareScale", and it tries to retrieve the batch-variations before the decoding process to fit the raw counts after decoding: specifically, a feedforward network is applied to the low-dimensional embedding and its corresponding batch embedding to produce a scaling vector the embedding. Then, the embedding is scaled by this scale vector and directed to the decoder. This modeling strategy delegates the extraction of transcriptional variation to the BBM, by placing the extraction of batch variations in the optimal timing or position: before and

after the extraction of transcriptional ones, effectively decouples these two variations in the multi-batch/multi-section scenario.

We now introduce the forward propagation in the combination of BBM and BEM to show the co-work of these two models. Technically, for a gene expression dataset $\boldsymbol{X}_b \in \mathbf{R}_+^{N\times G}$ encompassing N cells, G genes sequenced, and belongings to batch b in the batch indicator set B, we initialize a set of linear transformations $\mathbf{W} \in \mathbf{R}^{\mathrm{G}\times d\times \mathrm{K}} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \dots, \boldsymbol{W}_{\mathrm{K}}\}$ where d is the dimension of the final standardized embedding (default $d = 30$). This set is randomly initialized with a user-defined number of modules K (default $K = 32$) and a batch embedding $\boldsymbol{E}_B \in \mathbf{R}^{|B|\times D}$ where D is the dimension of non-standardized embedding. For any cell/spot $\mathbf{x}_{b,n} \in \boldsymbol{X}_b$ belonging to batch b, we perform the following transformation to map it to a sequence of gene modules: $\boldsymbol{M}_{b,n} = \mathrm{ReLU}\big(\mathrm{BatchAwareLayerNorm}\big(\mathbf{x}_{b,n}, \mathrm{onehot}(b)\boldsymbol{E}_B\big)\mathbf{W}\big)$.

Subsequently, we concatenate a randomly initialized "cls-token" $\mathbf{z}^{cls} \in \mathbf{R}^d$ to the sequence of gene modules to extract biological information and generate the final representation for each cell/spot. This concatenated sequence is then added by positional encodings $\boldsymbol{E}_{pos}$ and fed into 3 Transformer encoders. Afterwards, an extra feedforward layer named "Readout module" is used to derive the non-standardized low-dimensional representation, denoted as $\mathbf{h}_{b,n} \in \mathbf{R}^D$. This representation is then transformed to $\mathbf{z}_{b,n} \in \mathbf{R}^d$ using the reparameterization trick.

Before decoding $\mathbf{z}_{b,\mathrm{n}}$ back to original data space, the batch variations are retrieved for the alignment of reconstructions and raw gene expressions that indeed have batch-effects. Here, BatchAwareScale transforms the batch-independent embedding $\mathbf{z}_{b,\mathrm{n}}$ to a batch-dependent vector $\mathbf{z}'_{b,n} \in \mathbf{R}^d$.

As we model the gene expression data using a ZINB distribution, a non-linear decoder is employed to decode $\mathbf{z}'_{b,n}$ and to infer the parameters of the ZINB distribution for sampling the reconstructed data and calculating the negative log-likelihood (reconstruction error). The model training is based on the summation of the reconstruction error and the Kullback-Leibler (KL) divergence between the distribution of $\mathbf{z}_{b,n}$ and standard normal, scaled by a small coefficient β, thereby forming a β-VAE (Fig. 1a). The representation $\boldsymbol{Z}_b = \{\mathbf{z}_{b,n}\}_{n=1}^N$ is the final representation of $\boldsymbol{X}_b$, called standardized embedding.

Referring back to our earlier introduction, GCN in our spatial model (SpM) can be termed as a specialized filter. It operates on a graph constructed using spatial locations and BBM-extracted embeddings, emphasizing principal signals while minimizing non-dominant cell signals throughout the model training process. By training the model in this manner, wherein data reconstruction occurs from neighbor-aggregated embedding (Fig. 1a), a form of spatial regularization is achieved. Consequently, the resulting SRT embedding produced by STEP — denoted as $\widetilde{\boldsymbol{Z}} \in \mathbf{R}^{S\times d}$ (we use tilde to represent the spatially filtered or smoothed embeddings), where S locations are sequenced — exhibits robust performance in discerning spatial domains within tissues.

SpM is initially inserted between the calculations of Transformer encoders and reparameterization trick, and is trained by tuning a pre-trained BBM (and BEM for multi-section case). This involves two stages: 1. Training a BMM (and BEM); 2.

Training of the SpM together with the tuning of the BBM (and BEM). However, it is not necessary for a non-single-cell resolution SRT data to capture the sole transcriptional variations. As a result, in our pursuit of optimizing model performance, we implement a contrastive learning strategy akin to the SimCSE (Gao et al., 2021). This strategic approach (see Methods), termed as "fast mode", enhances the training process as an end-to-end style, accelerating the targeted extraction of spatial domain level heterogeneity.

After the training of STEP, we conduct the KMeans or Leiden algorithm on the embedding $\boldsymbol{Z} = ||_{b \in B} \boldsymbol{Z}_b$ for the clustering of cell types; and spatially refined $\widetilde{\boldsymbol{Z}} = ||_{b \in B} \widetilde{\boldsymbol{Z}}_b$, assuming the dataset is of SRT data, for the identification of spatial domains (Fig. 1b). In addition, for peeking cell type level heterogeneity in non-single-cell resolution SRT data, we integrate the workflows of two modalities by combining the annotated scRNA-seq data with the non-single-cell SRT data. We use just one BBM combined with BEM to map the data from both modalities into a shared space, termed as co-embedding. Subsequently, we train a decoder specific for the SRT to remap it back to the original data space and also a SpM for understanding the spatial context of SRT data. This integration enables us to comprehensively merge the data from both modalities while accomplishing tasks specific to each modality. Our target of this integrative analysis is to infer the cell type composition in each spot, referring as cell-type deconvolution, in more biological reasonable way: spots in the same spatial domain share the similar cell-type compositions. With this biological observation and the comparative representation obtained from the co-embedding stage, we build our embedding-based deconvolution algorithm with domain-wise and gene-expression-wise regularizations (Fig.1b and Methods).

In the next section, we will showcase the detailed demonstrations and superior performances of STEP with the combo of BBM and BEM in four atlas level scRNA-seq datasets and 2 simulation datasets, alongside with an additional single-cell resolution SRT dataset.

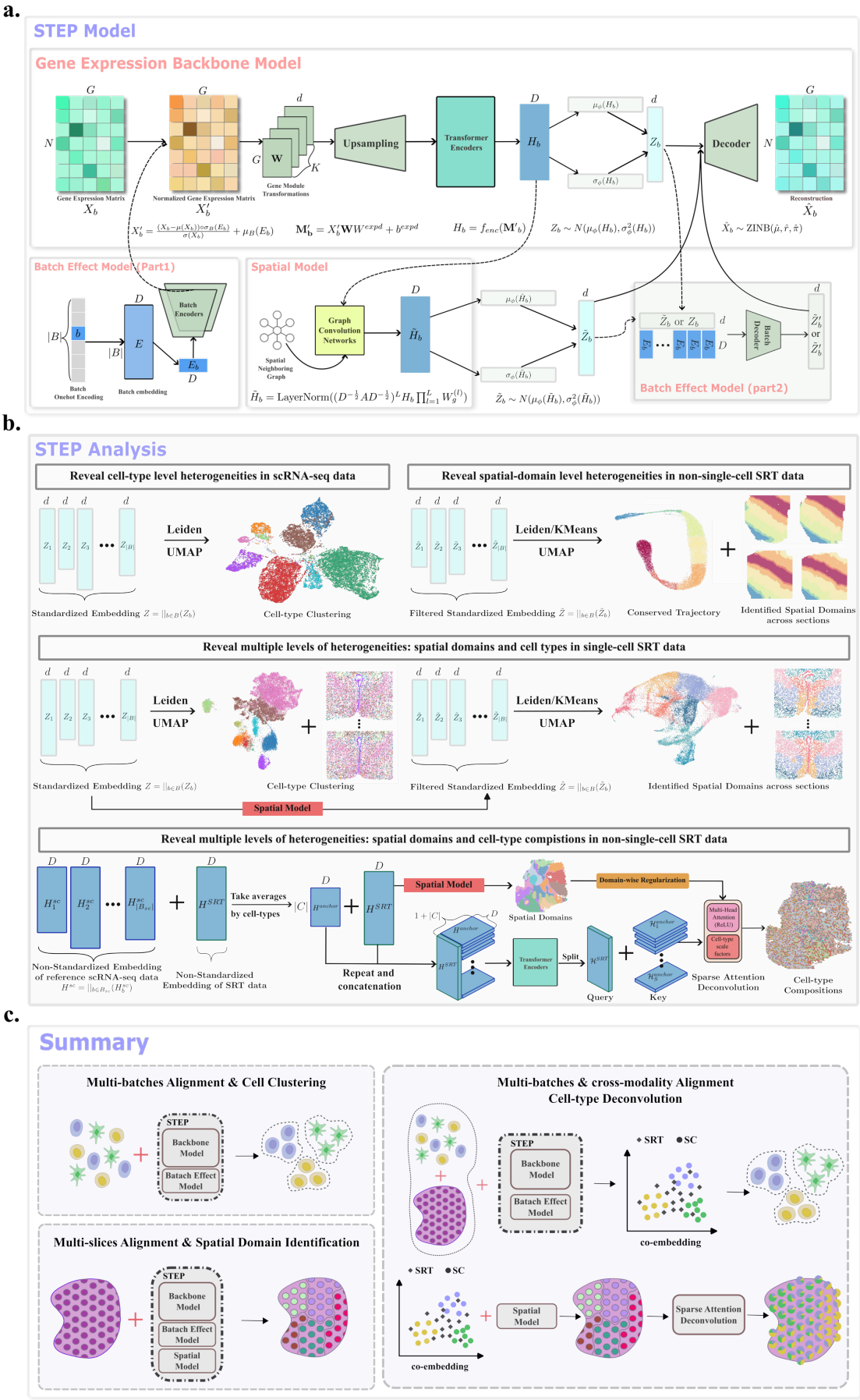
a.
STEP Model
Gene Expression Backbone Model
Gene Expression Matrix
Normalized Gene Expression Matrix
Gene Module Transformations
Upsampling
Transformer Encoders
Decoder
Reconstruction
Batch Effect Model (Part1)
Batch Encoders
Batch Onehot Encoding
Batch embedding
Spatial Model
Spatial Neighboring Graph
Graph Convolution Networks
Batch Decoder
Batch Effect Model (part2)
b.
STEP Analysis
Reveal cell-type level heterogeneities in scRNA-seq data
Leiden
UMAP
Standardized Embedding
Cell-type Clustering
Reveal spatial-domain level heterogeneities in non-single-cell SRT data
Leiden/KMeans
UMAP
Filtered Standardized Embedding
Conserved Trajectory
Identified Spatial Domains across sections
Reveal multiple levels of heterogeneities: spatial domains and cell types in single-cell SRT data
Leiden
UMAP
Standardized Embedding
Cell-type Clustering
Spatial Model
Leiden/KMeans
UMAP
Filtered Standardized Embedding
Identified Spatial Domains across sections
Reveal multiple levels of heterogeneities: spatial domains and cell-type compistions in non-single-cell SRT data
Non-Standardized Embedding of reference scRNA-seq data
Take averages by cell-types
Non-Standardized Embedding of SRT data
Repeat and concatenation
Spatial Model
Spatial Domains
Domain-wise Regularization
Transformer Encoders
Split
Query
Key
Multi-Head Attention (ReLU)
Cell-type scale factors
Sparse Attention Deconvolution
Cell-type Compositions
c.
Summary
Multi-batches Alignment & Cell Clustering
STEP
Backbone Model
Batach Effect Model
Multi-slices Alignment & Spatial Domain Identification
STEP
Backbone Model
Batach Effect Model
Spatial Model
Multi-batches & cross-modality Alignment
Cell-type Deconvolution
STEP
Backbone Model
Batach Effect Model
SRT
SC
co-embedding
Spatial Model
Sparse Attention Deconvolution

Fig. 1 Overview of STEP's architecture and workflows. a. The STEP framework consists of 3 main parts: backbone model, batch-effect model and spatial model and each part is tailored for different facts and issues encountered in the analysis of scRNA-seq data and SRT data; b. The compatible analyses with STEP's embeddings of different data modalities and resolutions; First, STEP can integrate multiple batches of scRNA-seq data integration and correctly uncover the different cell types; Next, for non-single SRT data, STEP can align multiple sections and identify spatial domains across sections; Also, when it comes to the single-cell resolution SRT data, STEP can reveal both cell types and spatial domains across the tissues. The last is the integration across modality that co-embeds the multi-batch SC data and multi-batch SRT data and achieve cell-type deconvolution for the non-single-cell resolution SRT data; c. The corresponding model combinations and settings for different analyses.

STEP integrates multiple batches of scRNA-seq data and single-cell resolution spatial transcriptomics by featuring adaptable batch-correction capabilities

We first evaluate STEP's BBM in two data, Human Colorectal Cancer (HCRC) and Mouse Small Intestine (MSI), from the newly proposed spatial transcriptomics technology called Visium HD by 10x Genomics (Genomics, 2024a, 2024b). This high-resolution spatial technology can resolve data at single-cell scale that consists of a grid with 2 x 2 μm barcoded areas, and the data is binned to 8 x 8 μm for analysis. The primary cell types and cellular components/structures are well identified by STEP's BBM with the inspection of the corresponding markers. For the HCRC data (bin size: 16 μm) STEP's BBM successfully deciphers cancer cells (cluster 0, 1, 6), cancer-associated fibroblasts (CAF) (cluster 3) and fibroblast cells (cluster 4) where the CAF cells surround and support the cancer cells as boundaries; goblet cells (cluster 2) surrounded by plasma cells (cluster 9) where they are co-locating in the intestine region and being walled by myofibroblasts cells (cluster 14); smooth muscle cells (cluster 7); endothelial cells (cluster 11) and epithelial cells (cluster 10) (Fig. 1a and Supplementary Fig.1). For the MSI data (bin size: 8 μm), BBM also manages to illustrate the sophisticated structure of intestine. Firstly, two cell clusters form an organizational protruding structure where absorption and immune functions are closely juxtaposed: nutrient-absorbing villus tip enterocytes (cluster 1, 2) envelop immune-active plasma cells (cluster 5). Notably, two kinds of villus tip enterocytes (cluster 1 and 2) are differentiated by the gradient expression profile of Reg3b (Supplementary Fig.2) along with the inner or outer segment of the intestine. Beneath, the intestine layer structure is also clearly revealed: the outermost layer, termed the "protective epithelial barrier layer" (cluster 4); the second layer, consisting of alternating clusters of "mucosal epithelial cells" (cluster 6) and "mixed immune-structural interface cells" (cluster 7); the middle layer, formed by "intestinal crypt secretory cells" (cluster 3); and the most innermost layer, termed "muscularis propria" and composed of smooth muscle cells (cluster 0). (Fig. 1b and Supplementary Fig.2). Another notable structure is lymphoid aggregates, including two big ones and a small one (cluster 8) where they are embedding between layers of intestinal crypt secretory cells and muscularis propria. Results of these two ultra-high resolution

and throughput data significantly indicate the robustness and effectiveness of BBM in capturing the transcriptional variations and deciphering the complex cellular structures in tissue microenvironment with the new sophisticated SRT technology.

Echoing our prior discussion, in the context of multi-batch single-cell resolution data integration, variations across different data batches manifest in two primary forms: batch variations and biological/transcriptional variations. Batch variations denote technical shifts in sequencing data distribution, whereas biological variations reflect the inherent heterogeneities among cells in distinct cell types or states. With the aim to eliminate batch variations while preserving biological variations, we proposed two solutions: 1. The combination of the BBM and BEM, referred as STEP for the simplicity. 2. STEP-anchored, where the BBM is further tuned by leveraging the pre-annotation when it is available. The latter method, STEP-anchored, operates under a semi-supervised learning paradigm: one "anchor point" is randomly initialized for each cell type as a cell-type-specific embedding and BBM is refined by introducing a classifier, replacing the decoder. This training process minimizes not only the KL-divergence but also the cross-entropy between embeddings of labeled cells and their respective labels, along with every anchor point and its corresponding label. This method combines pre-annotation results from all batches, enhancing the model's performance with a much more direct signal.

Evaluating data integration methods is a persistent challenge. Utilizing the comprehensive benchmarking strategy established by (Luecken et al., 2022), we manage to assess STEP across a spectrum of datasets, encompassing two simulation datasets and four real atlas-level datasets (including Lung atlas, human pancreas, human immune, human & mouse immune datasets), along with a single-cell resolution SRT dataset (Mouse hypothalamus data by MERFISH). This evaluation encompasses metrics for both batch effect removal and biological conservation. The overall assessment criterion combines the Bio Conservation score (weighted at 0.7) and the Batch Correction score (weighted at 0.3).

Remarkably, STEP-anchored, leveraging pre-annotation information, demonstrates superior performance across all datasets for both batch effect removal and biological conservation metrics (Fig.2c and Supplementary Table 1). Notably, STEP, without relying on pre-annotation information, achieves the highest score in the first simulation dataset and the human & mouse immune atlas dataset, outperforming other methods including the ones utilizing available cell type annotations.

It's important to highlight that within STEP-anchored, the "anchor points" efficiently anchor cell embeddings within their respective cell populations. This aspect notably bolsters the model performance in batch effect correction and biological conservation, although it may exhibit slightly less consistent trajectory conservation (Supplementary Fig.3-5). Conversely, STEP demonstrates a well-balanced performance in integration and trajectory conservation (Supplementary Fig.3-5). As a result, STEP-anchored shows promise as a tool for automated cell-type annotation, encompassing subtypes. Both STEP and STEP-anchored underscore the potency of our novel BEM combined with BBM, showcasing unmatched capabilities in capturing transcriptional

variations and eliminating non-biological noise.

Regarding multi-section single-cell resolution SRT dataset, specifically the mouse hypothalamus data by MERFISH (Moffitt et al., 2018), we select five consistent tissue sections from Bregma distances of -0.04 mm (5488 cells), -0.09 mm (5557 cells), -0.14 mm (5926 cells), -0.19 mm (5803 cells), -0.24 mm (5543 cells) from the original study's first animal. We utilize cell type annotations (15 cell types in total) from the original study as ground truth, excluding spatial location in this assessment. Comparing STEP against four other top-performing embedding-based integration methods (Fig.2a): scVI (Lopez et al., 2018), scANVI (Xu et al., 2021), Harmony (Korsunsky et al., 2019), and Scanorama (Hie et al., 2019), STEP surpasses all these methods. Additionally, we notice that a high value of principle components regression (PCR) often indicates an over-corrected state since the data itself does not appear to be affected by a severe batch-effect. Therefore, we exclude this metric from this test. In order to make up for this, we adjust the weights of Batch Correction score to 0.4 and Bio Conservation score to 0.6, respectively. In specific evaluation, STEP-anchored adeptly integrates the five sections (KBET: 0.82, Silhouette: 0.97), accurately defining cell types at the sub-type level through Leiden clustering (NMI: 0.85, ARI: 0.84). It achieves the highest scores for both Batch Correction (0.84) and Bio Conservation (0.67) with an aggregate score of 0.74 (Fig.2c). STEP without employing pre-annotation, closely follows STEP-anchored with a total score of 0.68 (Batch Correction 0.81; Bio Conservation 0.59; NMI 0.50; ARI 0.72), still outperforming the other four methods (Fig.2c). Evaluation of biological conservation reveals that both STEP-anchored and STEP excel in discovering and conserving both common and rare cell populations (Fig.2d). Furthermore, our two methods successfully delineate a boundary between Excitatory cells and Inhibitory cells in the UMAP (Fig.2e), a distinction that eludes the other methods.

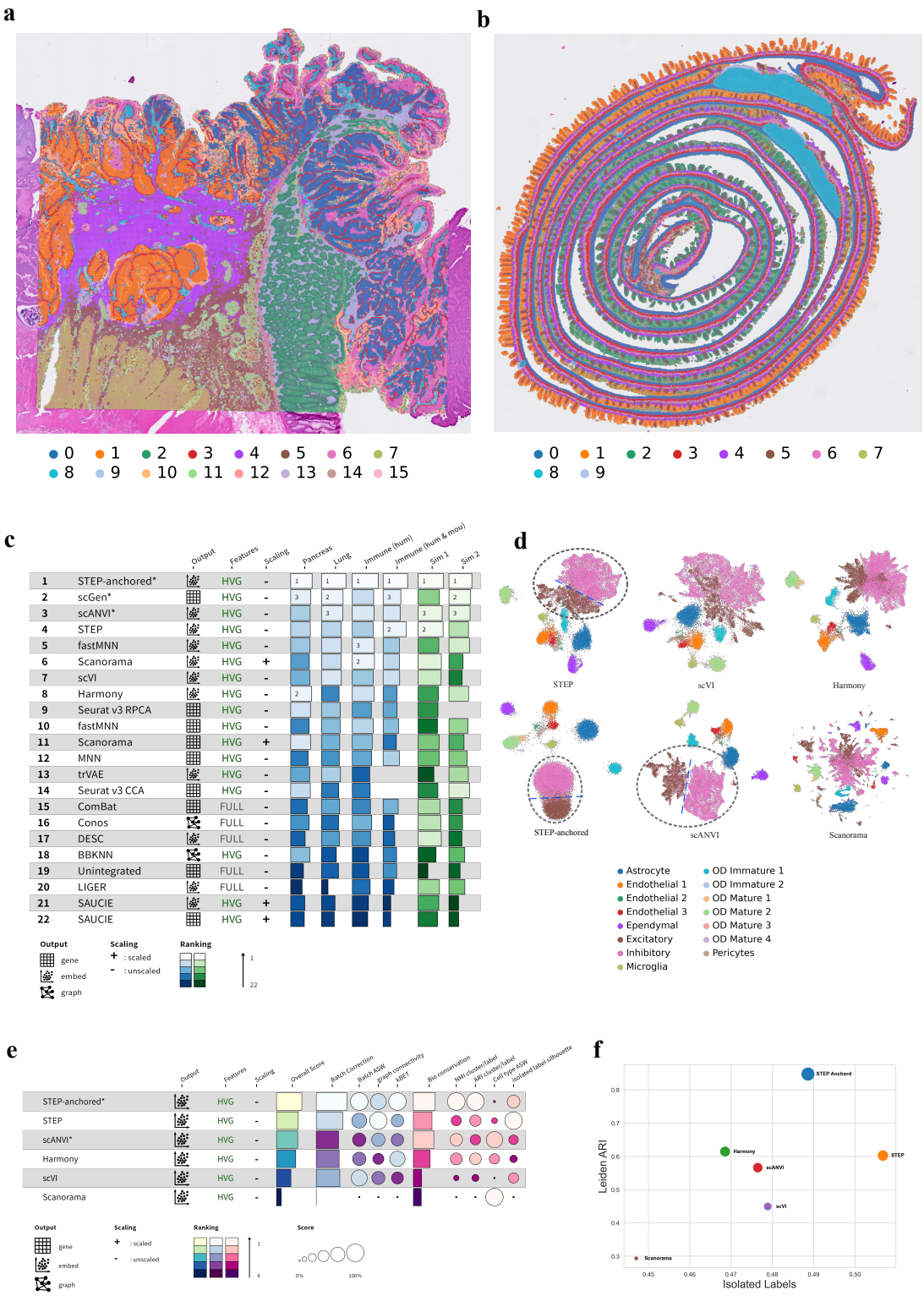

Fig. 2 Analyses of revelation of cell-type level heterogeneities in single-cell resolution datasets. a. Clustering results obtained from STEP for Human Colorectal Cancer (16 μm) by 10x Visium HD; **b**. Clustering results obtained from STEP for Mouse Intestine (8 μm) by 10x Visium HD; **c**. Ranking and bar plot for the performance of each tested methods with its the testing conditions (using HVGs only or all genes; scaled or not) and output style (graph, embedding or gene-expression). Tested datasets and results are obtained from scib, including 4 atlas-level scRNA-seq data by 10x Visium and 2 simulation datasets. The scores are weighted sum of Bio Conservation score (0.7) and Batch Correction score (0.3); **d**. UMAP plot of methods' results in MERFISH data, colored

by cell-type annotation from original study. The achievement of differentiating Excitatory and Inhibitory cells is represented by additional emphasis through circles with a dotted style and a straight dotted line; **e**. Ranking and bar plot for the performances of STEP, STEP-anchored, scANVI, Harmony, scVI and Scanorama in a MERFISH dataset; **f**. Bubble plot for how well the clustering results obtained from each tested method. The x-axis represents the isolated-label silhouette, and y-axis represents the ARI score for the clustering results using Leiden. The size of bubble is calculated by the 1000x of multiplication of the two metrics.

STEP accurately delineates spatial domains in human dorsolateral prefrontal cortex data in single-section scenario

For evaluating STEP's performance in single-section SRT data with the combination of BBM and SpM, we employ the widely used DLPFC dataset (Maynard et al., 2021) by Visium, encompassing 12 sections with manually annotated DLPFC layers and white matter (WM). Comparing STEP against GraphST (Long et al., 2023), STAGATE (Dong & Zhang, 2022), SpatialPCA (Shang & Zhou, 2022), BayesSpace (Zhao et al., 2021), and SpaGCN (Hu et al., 2021), we assess their proficiency in identifying spatial domains individually for each section. The accuracy of the identified spatial domains is quantified using the Adjusted Rand Index (ARI), measured against manual annotations (Fig.3a). Additionally, we gauge the quality of identified domains for spatial continuity and smoothness using Moran's I and Geary's C metrics. These metrics rely on spatial neighboring graphs constructed via a k-Nearest-Neighbor (kNN) approach with k values of 1, 10, 20, and 30.

Particularly focusing on the results of section 151673, STEP's identified domains are spatially smoother and more continuous compared to the other methods. GraphST, STAGATE, and SpatialPCA also exhibit smooth domains for several layers but fail to precisely capture the start of layer 2 as per manual annotations (Fig.3b). The results from all sections further showcase STEP's exceptional performance. Our method achieves the highest median ARI of 0.62 and an average ARI of 0.61 across all sections in single-batch analysis (Fig.3c, d and Supplementary Fig.6-8). Comparatively, GraphST, STAGATE, SpatialPCA, BayesSpace, and SpaGCN trail behind in median and mean ARI scores.

Quantitative measurements of spatial continuity via Moran's I and Geary's C further confirm STE's superior performances. Across all k values (1, 10, 20, 30), STEP consistently achieves the highest Moran's I scores and the lowest Geary's C scores, outperforming all the other methods across various k values (Fig.3e, f). Specifically, STEP maintains a mean Moran's I score of 0.98, exhibits its superiority in ensuring spatial continuity compared to other evaluated methods (GraphST: mean Moran's I = 0.86; STAGATE: mean Moran's I = 0.84; SpatialPCA: mean Moran's I = 0.72; BayesSpace: mean Moran's I = 0.75; SpaGCN: mean Moran's I = 0.73).

Beyond accurately identifying spatial domains, the underlying embeddings generated by STEP clearly reveal the tissue's developmental trajectory (Fig.3g and Supplementary Fig.9-11). Employing UMAP for dimension reduction to visualize the embeddings of each method, the results demonstrate that STEP maintains a distinctly striped

trajectory. Notably, STEP effectively delineates the WM/Layer 6 part. In contrast, STAGATE and spatialPCA illustrate the directional development relationship between domains but struggle to differentiate the WM part from other layers.

Further exploration of the embeddings and spatial domains derived from STEP involves in analyzing the autocorrelation of spatial embeddings. The heatmap of Pearson Correlation (Fig.3h) distinctly separates each domain, except for the WM and Layer 6, consistent with the UMAP visualization. Additionally, we conduct differential gene analysis between domains using the Wilcoxon rank sum test. This analysis successfully identifies several common spatially variable and domain-specific genes (Maynard et al., 2021), such as HOPX and HPCAL1 for layer 2, PCP4 for layer 3, and KRT17 for layer 4 (Fig.3j), further substantiating STEP's proficiency in spatial domain identification.

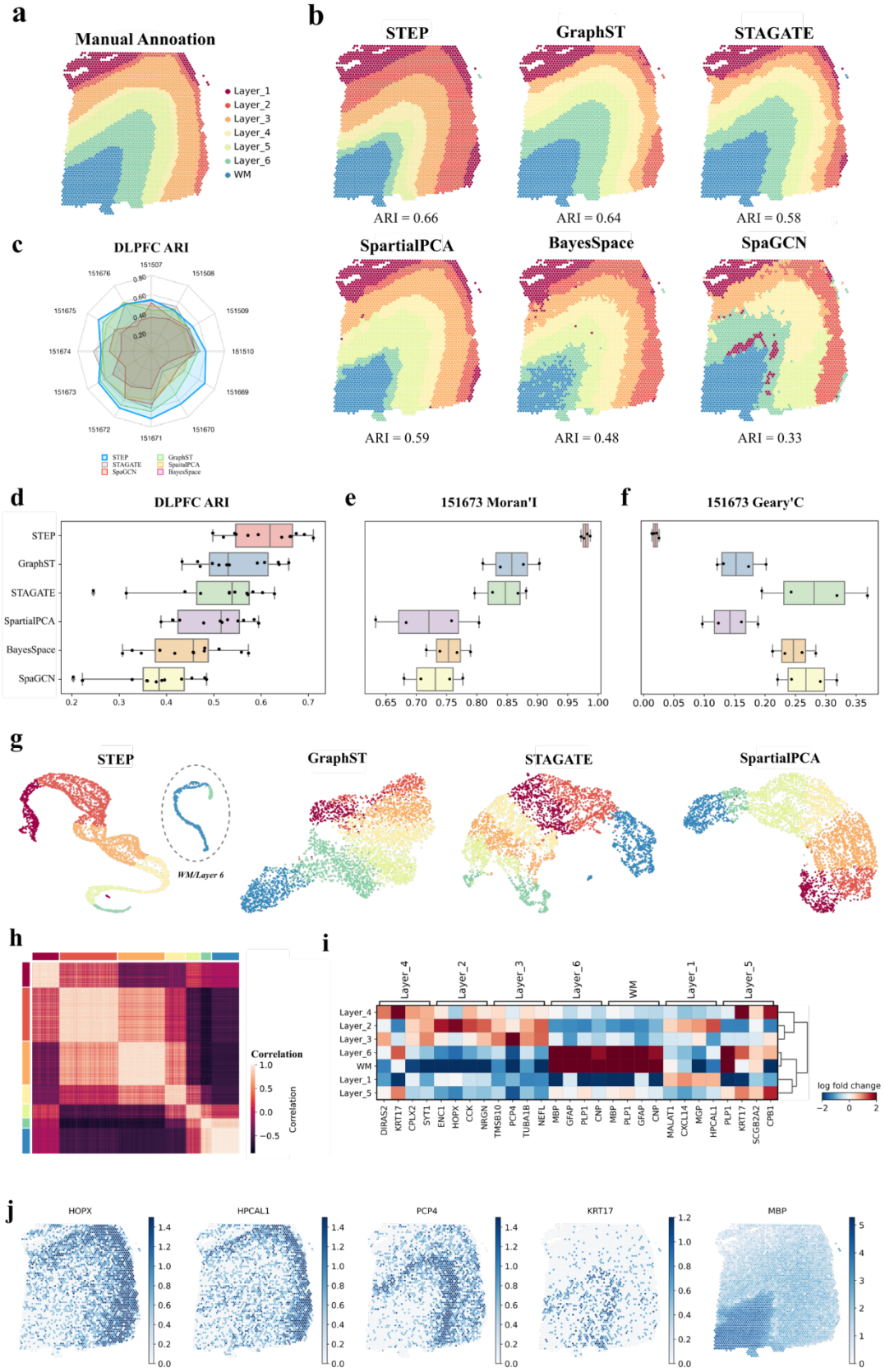

Fig. 3 Analysis of human DLPFC data at single-batch-wise. a. Manual annotation from original study for section 151673; **b**. Spatial domains identified by each tested methods, with name at the top and ARI score at the bottom; **c**. Radar plots of the ARI

scores for each method in each section; **d**. Box plots for the ARI scores of each method in each section; **e**, **f**. Box plots for Moran's I and Geary's C scores for the spatial domains in section 151673 obtained by each method; **g**. UMAP plots of embedding by STEP and compared methods; **h**. Heatmap of Pearson Correlation Coefficient by embeddings by STEP; **i**. Matrix plot of the log-fold-changes obtained by differentially expressed (DE) genes analysis based on STEP's identified spatial domains; **j**. Expression of domain-specific genes obtained by Spatially DE genes analysis.

STEP uniquely aligns diverse contiguous and non-contiguous tissue sections

A distinguishing feature of STEP is its capability to synchronize various tissue sections, even though they are non-contiguous, thus enabling a remarkable integration at the level of spatial domain heterogeneity. This feature corroborates to be especially beneficial when working with intricate biological structures, such as sections from the same tissue type but originating from different sub-regions or donors. This significant advantage stems from STEP's balanced modeling of transcriptional, batch and spatial variations with our modular design. By enabling a certain level of independence among these three models, STEP facilitates section-independent embeddings for SRT data, despite the SpM operating on a section-by-section basis. This inherent feature underscores the flexibility and resilience of STEP.

At the outset, we employ STEP with the combo of all three models (BBM, BEM and SpM) to integrate multiple sections in the DLPFC data, and we focus on the first 4 sections (P1): 151507 (4226 spots), 151508 (4384 spots), 151509 (4789 spots), 151510 (4634 spots), totaling 18033 spots, revealing a pattern akin to vertical alignment. Comparatively, we evaluate BASS-Mult (Li & Zhou, 2022), PRECAST (Liu et al., 2023), STAligner (Zhou et al., 2023) and MENDER (Yuan, 2024), four methods that integrate sections without dependence on the positional relationships between the sections. By visualizing uniformly identified spatial domains in each section, STEP consistently display ideal results akin to those observed in single-section analysis (Supplementary Fig.6-11). Except MENDER fails to identify most spatial domains, other three compared methods are all able to produce reasonable results at certain levels (Supplementary Fig.12). Quantitatively, STEP showcases similar ARI scores to those obtained through single-section analysis, outperforming all the other methods in every section (STEP: mean ARI = 0.53; BASS-Mult: mean ARI = 0.44; PRECAST: mean ARI = 0.32; STAligner mean ARI = 0.49) (Fig.4a). The spatial domains identified by STEP exhibit greater smoothness and continuity, supported by significantly higher and concentrated Moran's I scores (Fig.4b) and markedly lower and concentrated Geary's C scores (Supplementary Fig.13). Compared to STAligner and PRECAST, which are also embedding-based methods, a detailed examination of alignment results reveals that although both methods effectively remove batch effects, the trajectory from Layer 1 to Layer 6/WM is more accurately represented in the embedding by STEP than that by PRECAST (Fig.4c). The UMAP visualization indicates that the STAligner's and PRECAST's embeddings lack developmental structure concerning manual annotation,

underscoring the superior quality of embeddings produced by STEP. The other 2 groups in DLPFC dataset (P2: 151669 to 151672, P3: 151673 to 151676) are also tested. STEP yields an overwhelming performance on the p3 group as STEP completely resembles the layer 2 across 4 sections compared to manual annotations and achieves the highest ARI scores to date in every section (151673: 0.68; 151674: 0.67; 151675: 0.63; 151676: 0.63) in our knowledge (Supplementary Fig.12). The spatially DE gene analysis based on identified domains reveals the coherent gene expression pattern for the same domain across sections. For instance, genes including GFAP, HPCAL, ENC1, NEFL, and TMSB10 exhibit high expression levels in Layer1 to Layer5, respectively (Supplementary Fig.14, 15).

Subsequently, we investigate STEP's efficacy in integrating horizontally aligned sections using data from the Mouse Brain's sagittal anterior and posterior sections by 10x Visium (Genomics, 2020a, 2020b). Despite being divided for sequencing, these sections, originating from physically connected tissues, exhibit distinct spatial domains due to variations in functional brain areas (Fig.4d). The identification of 28 domains reveals a consistent appearance of domains across corresponding spatial locations referenced to the Allen Mouse Brain Atlas (Atlas): the layers of the cerebral cortex are seamlessly connected across the two sections. Furthermore, distinct sub-regions in regions like the hippocampal formation (HPF), hypothalamus (HY), and Thalamus (TH) are successfully reconstructed (Fig.4e). Embedding visualization illustrates five primary populations corresponding to MOB, CTX, CBX, HPF, and Caudoputamen (CP) regions (Fig.4e). Notably, STEP pinpoints unique regions in the sagittal anterior and posterior sections, such as the cerebellar cortex (CBX) and the main olfactory bulb (MOB), at a sub-region level: Vermal regions (VERM) with molecular layer (domain 2) and granular layer (domain 20), and fiber tracts (domain 12) for CBX region; granular layer (domain 13) and outer plexiform layer (domain 26) for MOB region (Fig.4f).

The examination of non-contiguous section integration is conducted using samples from mouse liver ATG7 KO autophagy, obtained via 10x Visium. Besides, as previously mentioned, the microenvironment in livers is particularly complex, thereby the desired spatial domains are more functional and less spatially continuous. The datasets in previous tests of STEP are most about brains, of which the microenvironment and structures are simpler and more well-organized. This liver dataset encompasses 4 sections, accounting for a total of 14531 spots where two are derived from wild-type mice, and the other two are from genetically modified or “knockout” mice (Fig.4g). STEP is tasked to identify 6 spatial domains, except 1 domain appearing to be capsule (Supplementary Fig.16, 17), successfully locating the desired Zone 1, Zone 2 and Zone 3 regions consistently across sections. Building on this foundation, STEP further excels by bringing to light the subtleties of liver architecture through the identification of subdomains—transitional areas that serve as interzones between the well-defined Zone 1 and Zone 2, and between Zone 2 and Zone 3. The revelation of these subdomains is particularly significant as it reflects the transitional gradients in gene expression and metabolic activity, offering a more nuanced view of hepatic zonation that echoes the dynamic nature of liver physiology (Fig.4h). Other four methods, BASS, PRECAST,

STAligner and MENDER are also tested on this liver zonation task. But they all fail to identify unified domains across all sections (Supplementary Fig.18), unable to handle the technical between sections or spatial biases within sections and the complexity of liver microenvironment. As we navigate the complex terrain of hepatic zonation, STEP stands out as a pioneering computational method, marking the first instance of an analytical tool that can delineate with accuracy the three primary zones of the liver.

To assess the biological accuracy of identified spatial domains and identify genes enriched in the periportal (PP) and pericentral (PC) regions, spatially differential expression (DE) gene analysis is performed (Supplementary Fig.19). The analysis unveils several genes that are wildly confirmed as marker genes of PP and PC regions (Ben-Moshe et al., 2019; Halpern et al., 2017; He et al., 2021; Wei et al., 2021), such as Cyp2f2, Cyp2e1, and Cyp2c29, significantly expressed in PP and PC (Fig.4i, j and Supplementary Fig.20).

These findings bear significant implications, showcasing STEP's robust capability to seamlessly integrate tissue sections irrespective of their relative spatial relationships. Furthermore, these results pave the way for exploring transcriptional differences together with spatial context through the integrative analysis of multiple SRT datasets.

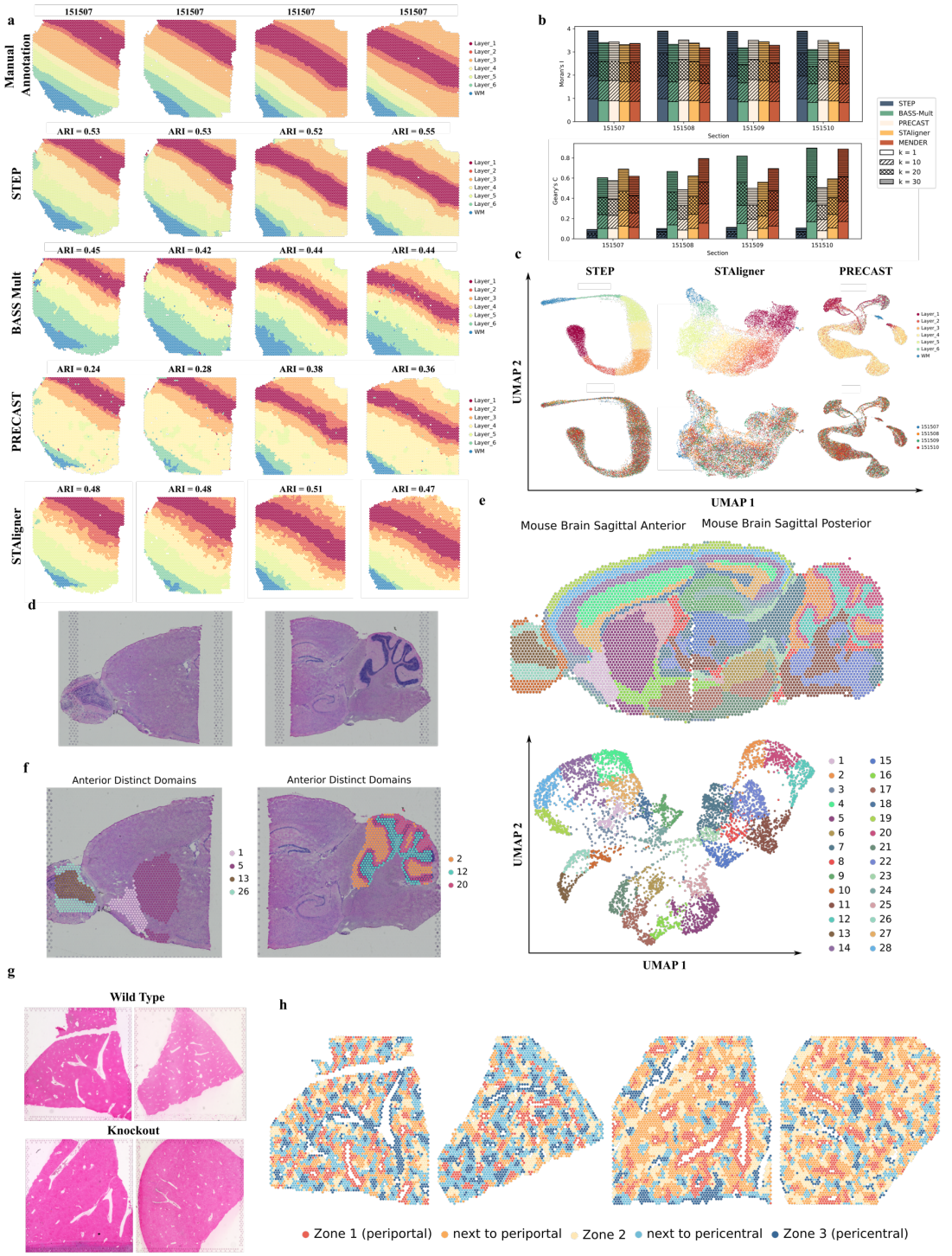
a
Manual Annotation
STEP
BASS Mult
PRECAST
STAligner
151507
ARI = 0.53
ARI = 0.52
ARI = 0.55
ARI = 0.45
ARI = 0.42
ARI = 0.44
ARI = 0.24
ARI = 0.28
ARI = 0.38
ARI = 0.36
ARI = 0.48
ARI = 0.51
ARI = 0.47
Layer_1
Layer_2
Layer_3
Layer_4
Layer_5
Layer_6
WM
b
Moran's I
Geary's C
Section
151507
151508
151509
151510
STEP
BASS-Mult
PRECAST
STAligner
MENDER
k = 1
k = 10
k = 20
k = 30
c
STEP
STAligner
PRECAST
UMAP 2
UMAP 1
d
e
Mouse Brain Sagittal Anterior
Mouse Brain Sagittal Posterior
f
Anterior Distinct Domains
Anterior Distinct Domains
1
5
13
26
2
12
20
UMAP 2
UMAP 1
g
Wild Type
Knockout
h
Zone 1 (periportal)
next to periportal
Zone 2
next to pericentral
Zone 3 (pericentral)
i
PP region
Cyp2f2
j
PC region
Cyp2c29

Fig. 4 Analysis of Multiple-section alignment and spatial domains identification for human DLPFC, Mouse Brain sagittal anterior and posterior and mouse liver autophagy ATG7 KO by 10x Visium. **a**. Manual annotation and spatial domains identified each tested method: STEP, BASS-Mult, PRECAST, and STAligner with corresponding ARI scores (MENDER fails to identify the desired spatial domains); **b**. Moran's I score for each method in each sample with each k value; **c**. UMAPs by embeddings obtained from STEP, STAligner and PRECAST, respectively, colored by spatial domain labels and section ids; **d**. Histology images for mouse brain sagittal anterior and posterior; **e**. Spatial domains identified by STEP across two sections (top panel) and UMAP colored by domain labels (bottom panel); **f**. Section-specific domains identified by STEP; **g**. Histology images of each section; **h**. Annotated spatial domains identified by STEP in each section: Zone 1, next to periportal, Zone 2, next pericentral and Zone 3; **i**. PP related domains and expressions of marker gene Cyp2f2; **j**. PC related domains and expressions of marker gene Cy2p2c29.

STEP demonstrates scalability in identifying domains within datasets featuring single-cell resolution in spatial transcriptomics.

While STEP has significantly enhanced our capacity to accurately cluster cell types within single-cell resolution SRT data in previous introductions, emphasizing the identification of spatial domains remains an essential objective in research. The paramount advantage of spatial domains, conceptively compared to cell types, lies in their reinforcement of our comprehension regarding how cells operate collectively within their specific spatial context. This insight is crucial for understanding tissue structure, disease progression, and the biological mechanisms underlying various cellular processes. To further evaluate STEP, we first conduct tests on Mouse hypothalamus data by MERFISH and Mouse medial prefrontal cortex data by STARmap, focusing on spatial domain identification in a multi-section setting.

The first MERFISH dataset utilized in this investigation comprises five consecutive tissue sections extracted from the preoptic region of the mouse hypothalamus. Across all tissue sections, meticulous cell annotations are conducted by (Li & Zhou, 2022), identifying eight distinct structures: the third ventricle (V3), bed nuclei of the stria terminalis (BST), columns of the fornix (fx), medial preoptic area (MPA), medial preoptic nucleus (MPN), periventricular hypothalamic nucleus (PV), paraventricular hypothalamic nucleus (PVH), and paraventricular nucleus of the thalamus (PVT).

Due to the difference of the spot/cell spatial alignment between 10x Visium and MERFISH that the spots of 10x Visium form a regular hexagonal grid while a transcript unit in MERFISH is generated from image segmentation, meaning there is no regular alignment pattern for these units. As a result, the neighboring graph is constructed as in k-Nearest-Neighbors (kNN) method based on transcript units' spatial locations. We set $k = 30$ (independent to the "k" in the evaluation of Moran's I and Geary's C) in kNN for this test, and again, STEP seamlessly eliminates batch-effects (Fig.5a). In comparison with manual annotations, STEP still generates smooth spatial domains with clear boundaries and the high-level matches (Fig.5b and Supplementary Fig.21), even the resolutions of transcripts units increase a lot. However, STEP mis-clusters the bottom corner area of MPA domain to fx, the upper area to MPN; and identifies more areas as PV than the its annotation across sections. Considering these domains are adjacent, this

may be caused by the relatively simple construction of the neighboring graph with kNN. BASS-Mult, as a method tailored for single-cell resolution SRT data, achieves similar performance with STEP, but better at distinguishing of these adjacent domains: MPA, MPN and fx in an overall view (Bregma -0.14, Bregma -0.19). PRECAST fails to identify any domains except the V3, this may be caused by the fact that the identification target of PRECAST is the minimal unit of the transcripts in each dataset. When the resolution comes to single-cell level rather than non-single-cell level, it still tends to capture the heterogeneity of cell types rather than spatial domains. STAligner and MENDER encounter the similar situation, they can only identify the V3 or near V3 regions but fails to decipher the other regions.

From a metrics perspective, STEP achieves ARI scores that are comparable across sections when matched against BASS-Mult, although slightly weaker in the first 4 sections (Fig.5c) (STEP: -0.04 ARI = 0.53, -0.09 ARI = 0.62, -0.14 ARI = 0.50, -0.19 ARI = 0.65, -0.24 ARI = 0.68; BASS-Mult: -0.04 ARI = 0.54, -0.09 ARI = 0.64, -0.14 ARI = 0.59, -0.19 ARI = 0.69, -0.24 ARI = 0.67). Contrastively, the smoothness of spatial domains by STEP is still dominant as the scores of Moran's I and Geary's C, outperforming other 2 methods in all cases of k values (Fig.5d) (average Moran's I by STEP: Bregma -0.04 = 0.89, Bregma -0.09 = 0.90, Bregma -0.14 = 0.90, Bregma -0.19 = 0.92, Bregma -0.24 = 0.93; average Moran's I by BASS-Mult: Bregma -0.04 = 0.77, Bregma -0.09 = 0.90, Bregma -0.14 = 0.82, Bregma -0.19 = 0.85, Bregma -0.24 = 0.83). We also conduct spatially DE gene analysis for the spatial domains by STEP to find domain-specific genes, and the results shows a reasonable overlap with original studies: Cd24a and Nnat1 enriched in V3 domain; Slc18a2 enriched in PV domain; Esr1 enriched in PV, MPN and BST domains; and MBP, Ermn enriched in MPA domain (Supplementary Fig.22).

Additionally, we also test STEP in the 7-th animal from the same original study with sections ranging wildly from Bregma distance 0.26 to -0.29, in total 12 sections (0.26, 0.21, 0.16, 0.11, 0.06, 0.01, -0.04, -0.09, -0.14, -0.19, -0.24, -0.29). This dataset is used to test if STEP can capture the dynamic changes of domains, not only the changes of domain types but also the changes of area for the same domain type, with the changes of distance and angles along the tissue. As a result, STEP effectively identifies anterior commissure (ACA) domain present at 0.26 to -0.14, as well as the area of ACA changes from disconnected to connected status (Supplementary Fig.23-26). And the identified V3 domain is at the bottom of the section at beginning, and it gradually "grows" up with the changes of Bregma distance along the tissue and reaches the top from -0.09 section (Supplementary Fig.25, 26), in the meantime, identified PV domain begins with a larger connected area and is gradually "squashed" by the V3 domain to edge-like area (Supplementary Fig.25, 26). The spatially DE results further support these developments (Supplementary Fig.27-29). This test strongly validates the capacity of STEP to handle with complex series of sections, even the data resolution is single-cell level.

Next, we test STEP, BASS-Mult and PRECAST on the Mouse medial prefrontal cortex data by STARmap. This series of sections (3 sections in total: BZ5, BZ9, BZ14)

consists of 4 domains including Layer1, Layer2/3, Layer5, Layer6. Besides testing the identification of spatial domains, we also test the robustness of embeddings when working with different clustering methods: Leiden and KMeans. Referring to the manual annotations, the spatial domains across all sections identified by STEP with two clustering methods show that they all, overall, work well with embeddings generated by STEP. This phenomenon further demonstrates the robustness and biologically meaningful features of these embeddings (Fig.5e and Supplementary Fig.31). However, STEP with Kmeans mis-clusters a set of cells in the Layer 6 domain as Layer 1 in section BZ9 (Fig.5e). Such mis-clustering situation can also be found in the domains obtained by BASS-Mult and PRECAST, but they mis-cluster not only this set of cells but also the cells in Layer 6 in all sections. In terms of quantity, the ARI scores (Fig.5f) show a decent matching between manual annotation and STEP with Leiden (BZ5: 0.88; BZ9: 0.80; BZ14: 0.86) and STEP with Kmeans (BZ5: 0.88; BZ9: 0.74; BZ14: 0.85) and competitive performances by BASS-Mult (BZ5: 0.87; BZ9: 0.66; BZ14: 0.77) and MENDER (BZ5: 0.77; BZ9: 0.80; BZ14: 0.85). PRECAST overall conserves the domain-level biological heterogeneity (BZ5: 0.49; BZ9: 0.42; BZ14: 0.53), but STAligner completely fails to identify the layer structure (Supplementary Fig.32). The significantly higher Moran's I and Geary's C scores with different k values again demonstrate the superiority in the purity or the spatial continuity of spatial domains generated by STEP (Fig.5g). We further compare STEP and PRECAST in the view of batch-effect correction measured by the "integration Local Inverse Simpson Index" (iLISI). Even though PRECAST achieves a higher iLICI score (2.17) than STEP (1.79), we can, by visualizing the manual annotations in the UMAPs, observe the significant over-correction by PRECAST and its loss of developmental trajectory (Fig.5h). Conversely, embeddings by STEP do not segregate apart but form a trajectory along the developmental orders of domains (Fig.5h). It is more reasonable since the spatial domains in this data, unlike the cell types, exhibit biological continuity because of the coarser graininess of biological heterogeneity and the nature of the development of the tissue. Additional spatially DE genes analysis successfully locates underlying genes that form the identified spatial domains (Supplementary Fig.32).

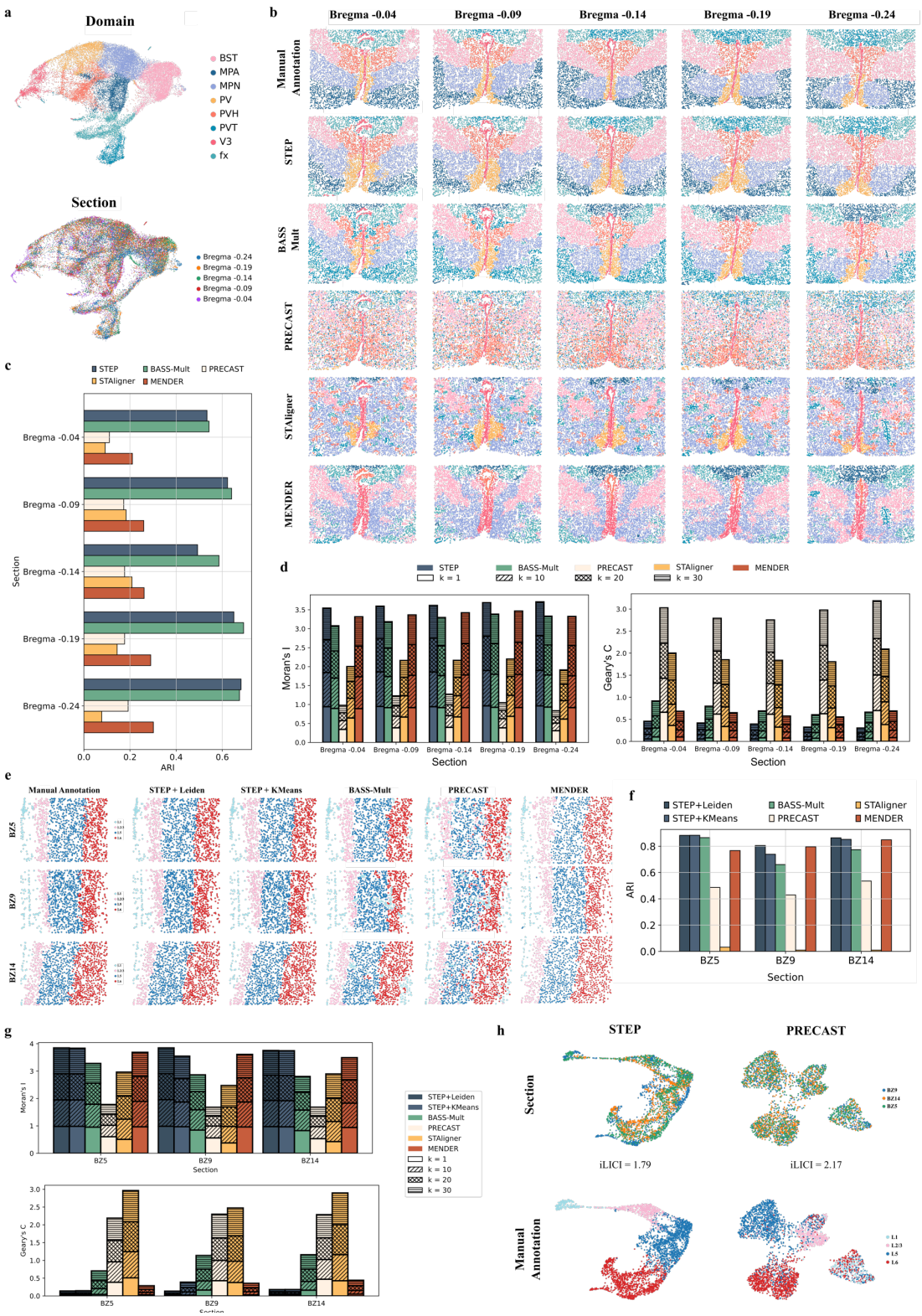

Fig. 5 Analysis of multi-section alignment and spatial domain identification in Mouse hypothalamus by MERFISH and Mouse medial prefrontal cortex by STARmap. a. UMAP plot obtained by STEP and colored by spatial domains and section ids; **b**. Comparison for the manual annotation and spatial domains identified by each test methods: STEP, BASS-Mult, PRECAST, STAligner and MENDER; **c**. Bar plots for ARI scores of each methods in each section by MERFISH; **d**. Stacked bar plots of Moran's I and Geary's C scores of each method in each section by MERFISH with each k value; **e**. Comparison for the manual annotation and spatial domains identified by each test methods: STEP+KMeans, STEP+Leiden, BASS-Mult, PRECAST,

STAligner and MENDER; **f**. Bar plots for ARI scores of each method in each section by STARmap; g. Stacked bar plot of Moran's I and Geary's C scores of each method in each section by STARmap with each k value; **h**. Comparison for the UMAP obtained by the embeddings of STEP and PRECAST, colored by manual annotation and section ids, respectively.

STEP seamlessly combines scRNA-seq data with spatial transcriptomics data, yielding intricate spatial domains and detailed deconvolution outcomes.

To conduct a joint analysis of scRNA-seq data and SRT data, STEP co-embeds both modalities of data into a shared latent space. In this approach, a domain is conceptualized as a potentially ambiguous cell type, suggesting that spots within the same domain may possess similar cell type compositions, referred as $\boldsymbol{P}_d \in \mathbf{R}^{|C|}$ for domain d. The representation of $\boldsymbol{P}_d$ as a random variable reflects the average cell-type composition within that domain. Instead of directly inferring the distribution and parameters of $\boldsymbol{P}_d$, STEP implicitly incorporates this information within the embedding of SRT data as a regularization term. This regularization term is computed as the sample variance of estimated cell-type composition $\boldsymbol{P}_d$ within the same domain. STEP primarily relies on the embedding and the resultant domains for cell-type deconvolution. However, it is notable that similar studies in image deconvolution have demonstrated that sampling from the latent space can lead to blurred reconstructions of the original data space images (Kingma & Welling, 2013), a phenomenon also observed in cell-type deconvolution studies (Lopez et al., 2022). To address this challenge, STEP integrates a reconstruction error directly derived from the gene expression level as an additional regularization technique in the deconvolution process. This dual-reconstruction approach separates the deconvolution problem into two sub-problems: identifying cell types occurrences and inferring coarse cell-type compositions in the spots through the latent space embedding, then refining the compositions in the gene-expression space (see Methods).

We initiate our deconvolution performance assessment with simulated data, labeled as pseudo-SRT. This simulation process involves a hierarchical strategy conducted on the reference scRNA-seq dataset (Fig.6a) and an actual tissue data (Fig.6b). The simulation commences by assigning an anticipated cell type composition (CTC) to each previously identified domain (Fig.6b). We refer such CTCs as domain-specific CTC. In this process, the CTC at each location is sampled from a normal distribution, using the mean equivalent to the corresponding domain-specific CTC. Moreover, we implement spatial smoothing on the sampled CTC for the simulation of spatial context. Next, each gene expression of pseudo-SRT is generated by convolving the sampled scRNA-seq data (gene expressions) with each CTC. The final step involves in multiplying each generated gene expression data by a randomly sampled cell number from a uniform distribution, with predefined lower and upper bounds of 10 and 20 (see Methods).

STEP is benchmarked against several other methods including CARD (Ma &

Zhou, 2022), Cell2location (Kleshchevnikov et al., 2022), RCTD (Cable et al., 2022), SPOTLight (Elosua-Bayes et al., 2021), Stereoscope (Andersson et al., 2020), DestVI (Lopez et al., 2022), and Seurat (Stuart & Satija, 2019) using simulation data. Metrics including Pearson correlation coefficient (PCC), Spearman correlation coefficient (SCC), Jensen–Shannon divergence (JSD), and mean squared root (MSE) are utilized for evaluation. Notably, STEP obtains competitive scores across all metrics: top-2 highest average PCC scores and SCC scores (STEP: PCC=0.93, SCC=0.76; RCTD: PCC=0.94, SCC=0.79; CARD: PCC=0.93, SCC=0.76; Cell2location: PCC=0.93, SCC=0.75; DestVI: PCC=0.88, SCC=0.73; Stereoscope: PCC=0.93, SCC=0.76; Seurat: PCC=0.71, SCC=0.42; SPOTlight: PCC=0.74, SCC=0.55); the 2 lowest average JSD scores (STEP: JSD=0.19; RCTD: JSD=0.20; CARD: JSD=0.17; Cell2location: JSD=0.21; DestVI: JSD=0.24; Stereoscope: JSD =0.19; Seurat: JSD =0.43; SPOTlight: JSD =0.31); and the top-3 lowest MSE scores (STEP: MSE=0.06; RCTD: MSE=0.06; CARD: MSE=0.06; Cell2location: MSE=0.05; DestVI: MSE=0.09; Steeoscope: MSE=0.06; Seurat: MSE=0.17; SPOTlight: MSE=0.13) (Fig.6c and Supplementary Fig.33, 34).

We also inspect the ability of each method to represent the spatial domains. Due to the inability of a direct inference for the spatial domains by the compared methods, we use two ways to infer the spatial domains from each CTC estimated by corresponding deconvolution methods: 1. Labeling the domain of each spot by the dominant cell types; 2. Labeling the domain of each spot by the argmax of the coefficients obtained by non-negative matrix factorization (NMF). The number of spatial domains labeled by the first way is not same with the ground truth due to the difference in the number of cell types and true spatial domains, while the latter labeling way yields the same number of spatial domains with ground truth by setting the number of components in NMF equal to the target number of spatial domains. ARI and NMI scores are then calculated to quantify the accuracy of inferred spatial domains. As a result, STEP achieves the highest scores in both metrics (ARI=0.70, NMI=0.77), outperforming all the other methods with all both two kinds of inferred spatial domains. The ARI scores of compared methods are all lower than 0.55 and the NMI scores are all lower than 0.74 (Fig.6d and Supplementary Fig.35).

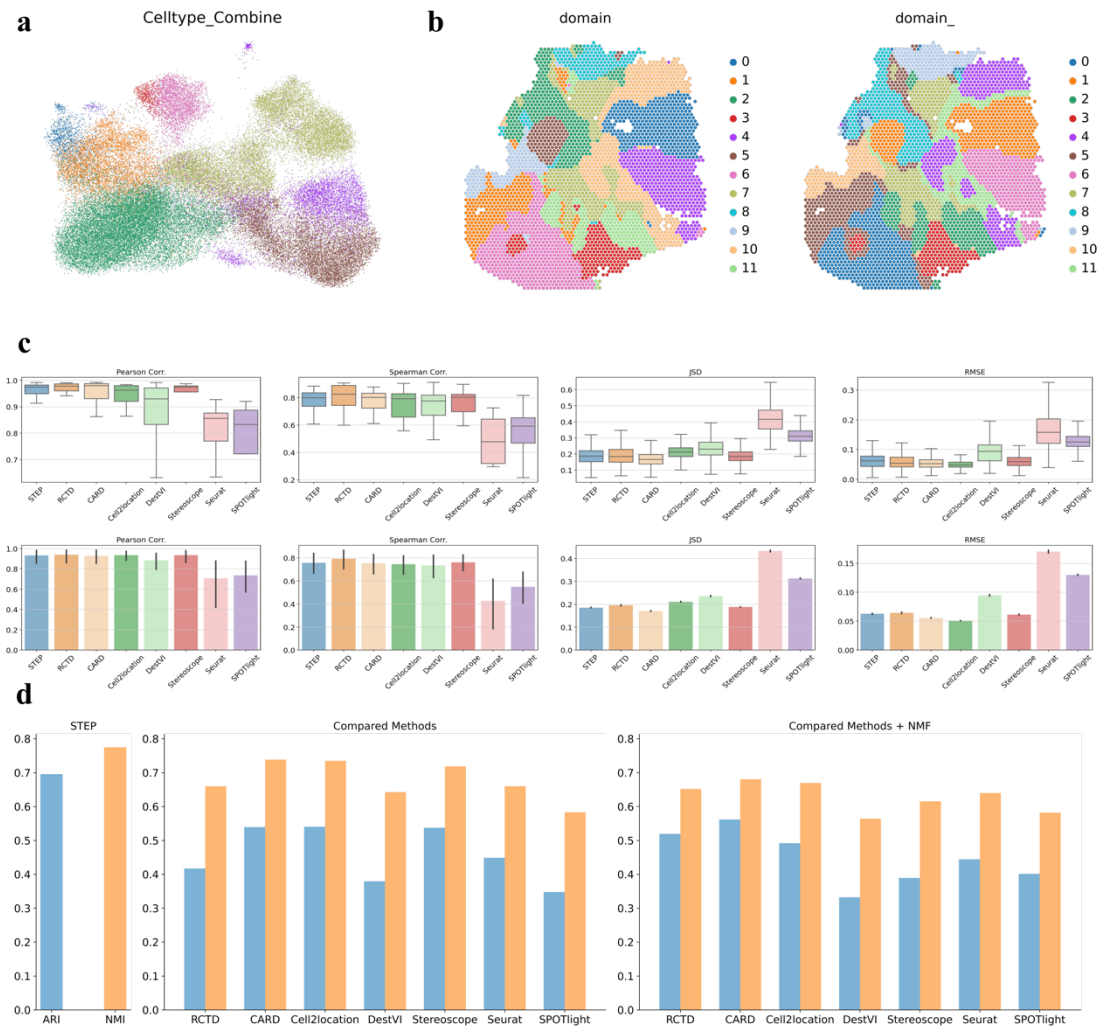

Fig. 6 Analysis of the spatial domain identification and deconvolution on simulated SRT data. a. UMAP plot for the reference scRNA-seq data; **b**. left: Ground truth for spatial domains, right: identified domains by STEP; **c**. Comparison for the accuracy of deconvolution through 4 metrics: PCC, SCC, JSD and RMSE (top: Boxplot, bottom: Bar plot); **d**. ARI and NMI scores for inferred spatial domains from each method.

STEP decrypts liver zonation and cellular compositions

Refocusing our attention on the intricate landscape of liver zonation, we find that within this realm, STEP exhibits the groundbreaking capabilities aforementioned. The precision in identifying Zones 1, 2, and 3 serves as a testament to the method's robustness and is also pivotal in advancing our understanding of liver function and pathology. Thanks to STEP's ability to seamlessly integrate multiple sections independent of spatial sources, unified spatial domains across these sections can be identified. Capitalizing on this feature, we proceed with a downstream analysis based on the delineated and annotated zones. This enables us to explore the biological differences between identical zones in WT and KO sections.

Starting with a summary of the zonation (Fig.7a), a high consistency of zone compositions is observed within the KO sections. There's a slightly lower consistency within the WT sections, likely due to the significantly fewer sequenced spots in the A-

WT section. This contrast further illustrates the robustness of STEP in handling imbalanced sample sizes. Subsequently, by comparing the zone compositions between WT and KO sections, it shows the significant increase of Zone 2 and the transition zone of Zone 1 (next to periportal), and the decrease of Zone 3 with its transition zone (next to pericentral). To investigate the transcriptomic reasons of the changes in zone compositions, we perform differentially expressed genes (DEG) analysis between WT and KO sections on each domain. The DEG results reveal remarkable discrepancies in the expression of key marker genes across varying conditions: a reduction of Spp1 and Clu in WT sections, along with a decrease in Cyp3a11 and Cyp2e1 in KO sections, offering compelling transcriptomic evidences to elucidate the expansion and contraction of the corresponding zones post-knockout (Fig.7b).

We simultaneously run a gene set enrichment analysis (GSEA) to inspect the cellular functions in each domain. For the three main zones (Zone 1, Zon2, and Zone 3), terms including GPCR, G protein, Synapse transmission, Vision, Immune [T cell, NK], Cardiac function, transduction, Cell interactions/cell membrane channels/integrin, Muscle contraction, Cilium move/structure, Extracellular matrix/components are found related to Zone 1; terms including Innate/adaptive immune, Immune activation, Cell activation [T cell, Lymphocyte], Cytokine production, Immune system activation signal transduction between cells, Neutrophil, Microglia phagocytosis, Cell interaction, Leukocyte production/migration/activation, Extracellular structure/integrin/collagen formation, Cell division/cycle, Mitotic, DNA repair are related to Zone 2; terms Innate/adaptive immune, Immune activation, Cell activation, Cytokine production, Immune system activation signal transduction between cells, Neutrophil, Microglia phagocytosis, Cell interaction, Leukocyte production/migration/activation, Extracellular structure/ integrin/ collagen formation, Cell division/cycle, Mitotic, DNA repair are found downregulated in the Zone 3 region. For the transition zones, the GSEA result shows "next to periportal" zone is active in division and migration with terms Cell projection/migration, Cell division/mitotic, Cell cycle, Chromosome structure, mRNA/DNA repair, Organelle formation/maintenance/location being significantly correlated; terms Mitochondrial matrix/protein, Catabolic, Metabolic, Synthesis, Aerobic respiration/Oxidative phosphorylation/Energy production, mRNA Processing, peroxisome are found correlated to the "next pericentral zone" (Supplementary Fig. 36-38).

Additionally, the pseudotime analysis is conducted for the further research on the development relationship between zones (Fig.7c). Upon visualizing the trajectories spatially, we notice the trajectories tend to originate from Zone 1 and its transition zone. With this in mind, we further examine the source zones and the target zones in every trajectory (Fig.7d). The summary of trajectories' directions explicitly supports our observation, and reveal a unified transition process between zones across sections: Zone 1 as the very beginning can develop to any other zones; "next periportal" zone, highly similar with Zone 1, can develop to any other zone except Zone 1; Zone 2 as the third stage can transition to "next to pericentral" zone, Zone 3 and itself; "next to pericentral" zone is the fourth stage in transition and can almost only develop to Zone 3, the last stage in the transition process. Given this lucid development lineage, we try to

differentiate the transition status of WT and KO. Taking the development of WT as a standard, KO's development appears to be unbalanced or immature, with the chain Zone 1 -> "next to periportal" zone -> Zone 2 being overactive, along with the inactivity in chain Zone 1 -> "next to pericentral" -> Zone 3 (Fig.7e).

In order to delve deeper into the cellular functions within each zone, we use STEP to perform cell type deconvolution using scRNA-seq data from (Guilliams et al., 2022) as reference. This approach allows STEP to dissect multi-scale biological heterogeneities, a capacity that not only enables the precise differentiation of liver zones, but also uncovers the inherent complexities within different cell types. The deconvolution results show that the area where hepatocytes appeared is significantly reduced in the KO sections, especially in the Zone 2 and Zone 3 regions. And we can observe that Central Vein Endothelial cells (CVE), are significantly reduced in KO samples. In contrast, the average proportion and the area of occurrence of Stellate cells increases significantly (Fig.7f, g and Supplementary Fig.39). Gathering deconvolution results in all sections, a high consistency of cell type compositions can be observed within KO sections and some domain-specific cell types are discovered, like Central Vein Endothelial cells, Lymphatic Endothelial cells and Stellate cells (Fig.7h).

With STEP's proficiency in identifying these zones and interzones, we gain a comprehensive view that not only reinforces the established understanding of hepatic biology but also extends it by adding depth to the spatial continuum of liver functionality. Such detailed zonal mapping is of immense value in the study of hepatic diseases, where understanding the zonation can lead to insights into zonal susceptibility, disease progression, and potential therapeutic interventions.

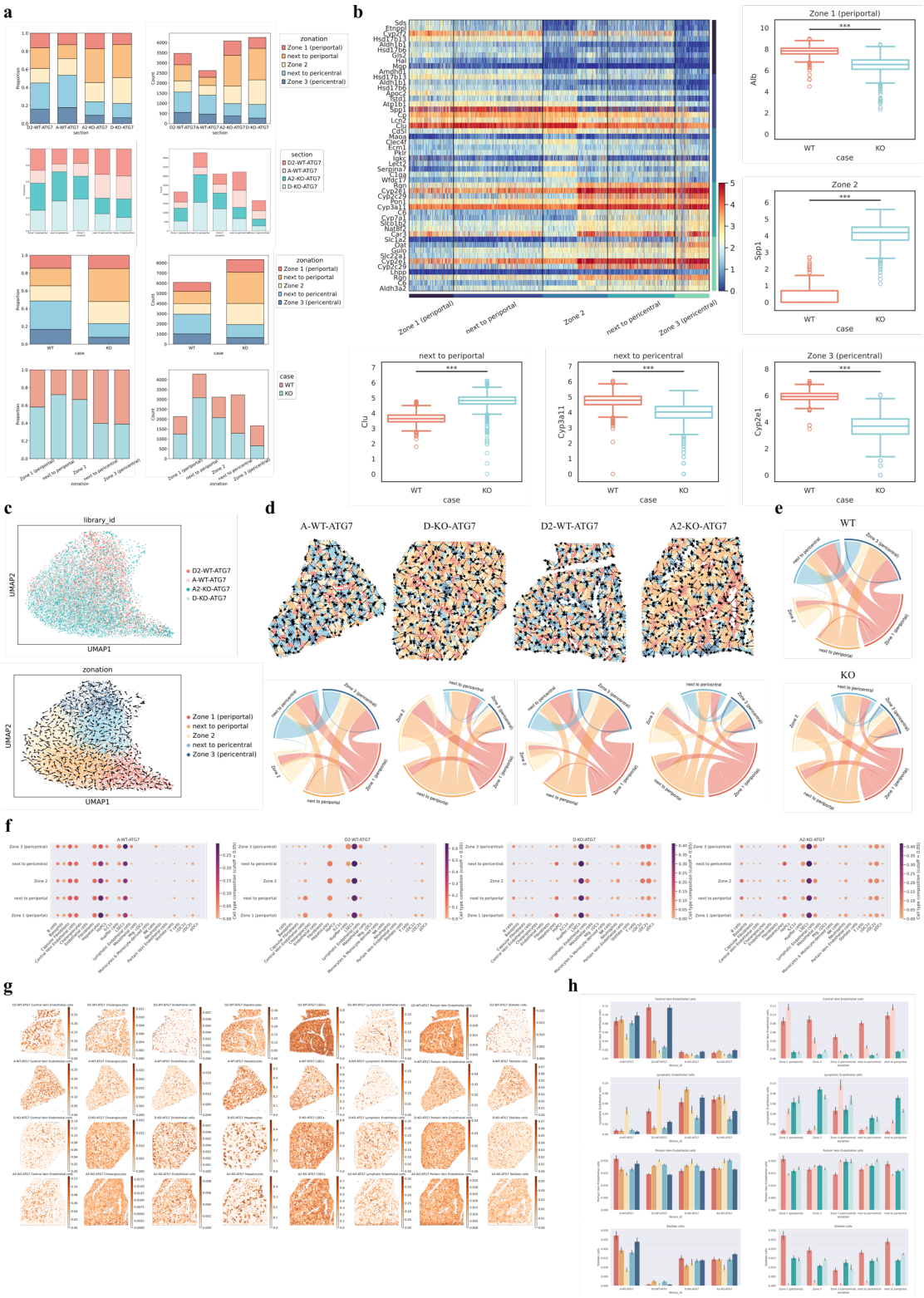

Fig. 7 Downstream analysis of the identified spatial domains and cell type deconvolution analysis on Mouse Liver Atg7 KO Autophagy data. a. Zonation summary: left column denotes proportion; right column signifies count. First row gives proportion and count of total spots in each domain per section; second row, the spots' proportion and count from each section in each domain; third row, proportion and count of total spots in each domain per condition; fourth row, proportion and count of spots from each condition in each domain; **b**. Heatmap of top DEGs between spatial domains identified and boxplots of DEGs between WT and KO in each domain; **c**. UMAP produced from smoothed embedding by STEP, color-coded by section IDs (top) and zonation

(bottom); **d**. Top: spatial plot indicating arrows from lowest pseudotime value in each grid to the highest one. Bottom: chord diagram of arrow directions between each domain pair in every section; **e**. Chord diagram of directions of arrows between each pair of domains in each condition; **f**. Cell type deconvolution results showing cell type composition (CSC) cut-off at 0.05 and normalized along cells. Dot size exhibits the proportion of total spots with corresponding non-zero CSC values; **g**. Spatial plot of CSC values of selected cell type; h. Bar plots of CSC distributions per domain per section (left column) and per section per domain (right column).

Discussion

The proposed foundation AI architecture for SRT data, STEP, effectively decouples the modeling and gene-expression, batch-effect and spatial correlation, and balances these variations. In this way, STEP exhibits the capacity on the integration of multiple batches/sections and identification of different grained heterogeneities (cell types and spatial domains) and generalization to different resolution of transcriptomics data. STEP also successfully bridges the gap between spatial domain identification and cell-type deconvolution, providing a unified framework to decode the intricate biological systems better. This is a notable deviation from most current analytical solutions that primarily fail to balance the different sources of variation.

In more details, the novelty of STEP's backbone model lies in its ability to model gene expression data using learnable gene module transformations together with Transformer, allowing a much more complex capture of transcriptional variations. And using GCN as a spatial filter in STEP's spatial model enables fast and effective identification of spatial domains by preserving the principal biological signals in a cell or spot. Contrasting to existing graph neural networks (GNNs) based methods like GraphST, STAGATE and SpaGCN, STEP not only performs better identification of spatial domains, but also demonstrates better consistency in model utilization with the definition of spatial domains. Moreover, our newly proposed paired batch-effect model further enhances the robustness of STEP when facing to multiple batches or sections, even in the cross-modality situation. This elaborate architecture enables STEP to fully capitalize on interaction within gene modules and spatial information within the data and can be easily generalized to multiple batches scenario, a facet lacking in current techniques. Our approaches, therefore, contribute to a more nuanced understanding of both non-single-cell and single-cell resolution spatial transcriptomics data. Of considerable significance is STEP's performance across multiple tasks, including cell types identification and data integration in various batches, platforms and protocols; spatial domains identification and integration of both contiguous and non-contiguous sections; and cell-type deconvolution in integrative analysis across modalities. STEP exhibits a notable proficiency in dealing with varying complexities of dataset management, showcasing a higher level of accuracy compared to the existing solutions.

Despite promising results, there are areas that require further investigation. For instance, embeddings are not fully utilized in cell-type deconvolution, the gene expression imputation performance is not as remarkable as other tailored methods, and our method relies on other selection methods for highly variable genes (HVGs) (Lause et

al., 2021; Satija et al., 2015; Zheng et al., 2017) or spatially variable genes (SVGs) (Zhu et al., 2021). These constraints primarily stem from a lack of a more powerful decoder and a superior pretext task beyond simple gene-expression profile reconstruction. Similarly, the potential inconsistency in STEP's encoding and decoding processes is also evident. While the encoding part is designed to model underlying biological effects, the decoding part is solely focused on modeling the statistical properties of observed gene expressions. Ideally, the decoding process should also model gene modules' interactions, even though this might intensify the computational load. Furthermore, there are methods based on the Transformer and trained within a self-supervised paradigm (Cui et al., 2024; Yang et al., 2022). However, these methods often mirror the "Masked Language Model" (Devlin et al., 2018; Floridi & Chiriatti, 2020), a prevalent training paradigm in natural language processing (NLP). In this paradigm, the expressions of some genes are randomly masked during training, and the model is encouraged to predict the masked values by "referring" to other genes' expressions. These methods overlook the fact that genes operate in groups as gene modules and assume uniform importance across all genes. Consequently, despite being a good attempt, these methods might be further improved by incorporating our biological modeling principals. There's another method for cell-type annotation called TOSICA (Chen et al., 2023) which also incorporates the idea of interactions of gene modules (the term "gene-set" is used in their work, instead). Nevertheless, TOSICA uses the predefined gene-set from pathway as exerting knowledge, and this may restrict the exploration of subtle and dynamic changes in the biological activities. Conversely, our intent is to develop an explainer based on STEP to better interpret the biological activities captured by STEP, such as gene regulation networks and pathways. In addition, hyper-parameter tuning also poses challenges, a native problem for deep learning models that may comparatively hinders STEP to reach the potentially highest usability and optimal performance. Going forward, we plan to employ strategies like model distilling to simplify STEP's training or tuning processes and increase its consistency. Our future focus is to address these limitations and refine the model in upcoming works.

Additionally, while most of our experiments focus on 10x Chromium, 10x Visium, MERFISH, STARmap protocols, exploring STEP's applicability on other datasets can contribute to its further refinement. The potential integration of STEP with other analytical methodologies such as Slide-seq V2 and Stereo-seq, might also lead to unforeseen synergistic effects, further enriching our understanding of spatial transcriptomics and single-cell RNA sequencing. Taking note of recent works around the development of more complex and significantly larger spatially resolved single-cell atlases (Greenbaum et al., 2023; Hickey et al., 2023; Lake et al., 2023; Zhang et al., 2023), we aim to apply STEP/STEP-anchored on such datasets, and such possibilities open exciting avenues for the future research.

In essence, our work forms an initial yet crucial step towards the identification of different types of biological heterogeneity with integration of multiple batches and sections in the meantime. Also, our work demonstrates a competitive integrative analysis

across two modalities and underscores the need for more robust analytical strategies and tools in the field.

References

Andersson, A., Bergenstrahle, J., Asp, M., Bergenstrahle, L., Jurek, A., Fernandez Navarro, J., & Lundeberg, J. (2020). Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun Biol*, *3*(1), 565. https://doi.org/10.1038/s42003-020-01247-y

Atlas, A. R. *Allen Reference Atlas – Mouse Brain*. http://atlas.brain-map.org/

Ben-Moshe, S., Shapira, Y., Moor, A. E., Manco, R., Veg, T., Bahar Halpern, K., & Itzkovitz, S. (2019). Spatial sorting enables comprehensive characterization of l iver zonation. *Nat Metab*, *1*(9), 899-911. https://doi.org/10.1038/s42255-019-0109-9

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., & L erchner, A. (2018). Understanding disentangling in β-VAE. https://doi.org/10.48550/arXiv.1804.03599

Buttner, M., Miao, Z., Wolf, F. A., Teichmann, S. A., & Theis, F. J. (2019). A te st metric for assessing single-cell RNA-seq batch correction. *Nat Methods*, *16*(1), 43-49. https://doi.org/10.1038/s41592-018-0254-1

Cable, D. M., Murray, E., Zou, L. S., Goeva, A., Macosko, E. Z., Chen, F., & Iri zarry, R. A. (2022). Robust decomposition of cell type mixtures in spatial transcri ptomics. *Nat Biotechnol*, *40*(4), 517-526. https://doi.org/10.1038/s41587-021-00830-w

Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., & Han, J. J. (2023). Transformer f or one stop interpretable cell type annotation. *Nat Commun*, *14*(1), 223. https://doi.org/10.1038/s41467-023-35923-4

Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S., & Zhuang, X. (2015). RN A imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *S cience*, *348*(6233), aaa6090. https://doi.org/10.1126/science.aaa6090

Chidester, B., Zhou, T., Alam, S., & Ma, J. (2023). SPICEMIX enables integrativ e single-cell spatial modeling of cell identity. *Nat Genet*, *55*(1), 78-88. https://doi.org/10.1038/s41588-022-01256-z

Clifton, K., Anant, M., Aihara, G., Atta, L., Aimiuwu, O. K., Kebschull, J. M., Miller, M. I., Tward, D., & Fan, J. (2023). STalign: Alignment of spatial transcrip tomics data using diffeomorphic metric mapping. *Nat Commun*, *14*(1), 8123. https://doi.org/10.1038/s41467-023-43915-7

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., & Wang, B. (2024). s cGPT: toward building a foundation model for single-cell multi-omics using gener ative AI. *Nat Methods*. https://doi.org/10.1038/s41592-024-02201-0

Devlin, J., Chang, M.-W., Lee, K., Toutanova, K., Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transform

ers for Language Understanding. *arXiv e-prints*. https://doi.org/10.48550/arXiv.1810.04805
Dong, K., & Zhang, S. (2022). Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat Commun*, *13*(1), 1739. https://doi.org/10.1038/s41467-022-29439-6
Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I., & Heyn, H. (2021). SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res*, *49*(9), e50. https://doi.org/10.1093/nar/gkab043
Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. Proceedings of the IEEE/CVF international conference on computer vision,
Floridi, L., & Chiriatti, M. (2020). GPT-3: Its Nature, Scope, Limits, and Consequences. *Minds and Machines*, *30*(4), 681-694. https://doi.org/10.1007/s11023-020-09548-1
Gao, T., Yao, X., Chen, D., Gao, T., Yao, X., & Chen, D. (2021). SimCSE: Simple Contrastive Learning of Sentence Embeddings. *arXiv e-prints*. https://doi.org/10.48550/arXiv.2104.08821
Gao, Z., Cao, K., & Wan, L. (2024). Graspot: A graph attention network for spatial transcriptomics data integration with optimal transport. *bioRxiv*. https://doi.org/10.1101/2024.02.01.578505
Genomics, x. (2020a). *Mouse Brain Serial Section 1 (Sagittal-Anterior), Spatial Gene Expression Dataset* Version Space Ranger 1.1.0). https://www.10xgenomics.com/cn/datasets/mouse-brain-serial-section-1-sagittal-anterior-1-standard-1-1-0
Genomics, x. (2020b). *Mouse Brain Serial Section 1 (Sagittal-Posterior), Spatial Gene Expression Dataset* Version Space Ranger 1.1.0). https://www.10xgenomics.com/cn/datasets/mouse-brain-serial-section-1-sagittal-posterior-1-standard-1-1-0
Genomics, x. (2024a). *Human Colorectal Cancer (FFPE), HD Spatial Gene Expression Dataset* Version Space Ranger 3.0.0). https://www.10xgenomics.com/cn/datasets/visium-hd-cytassist-gene-expression-libraries-of-human-crc
Genomics, x. (2024b). *Mouse Small Intestine (FFPE), HD Spatial Gene Expression Dataset* Version Space Ranger 3.0.0). https://www.10xgenomics.com/cn/datasets/visium-hd-cytassist-gene-expression-libraries-of-mouse-intestine
Greenbaum, S., Averbukh, I., Soon, E., Rizzuto, G., Baranski, A., Greenwald, N. F., Kagel, A., Bosse, M., Jaswa, E. G., Khair, Z., Kwok, S., Warshawsky, S., Piyadasa, H., Goldston, M., Spence, A., Miller, G., Schwartz, M., Graf, W., Van Valen, D., . . . Angelo, M. (2023). A spatially resolved timeline of the human maternal-fetal interface. *Nature*, *619*(7970), 595-605. https://doi.org/10.1038/s41586-023-06298-9
Guilliams, M., Bonnardel, J., Haest, B., Vanderborght, B., Wagner, C., Remmerie, A., Bujko, A., Martens, L., Thone, T., Browaeys, R., De Ponti, F. F., Vanneste, B., Zwicker, C., Svedberg, F. R., Vanhalewyn, T., Goncalves, A., Lippens, S., Devriendt, B., Cox, E., . . . Scott, C. L. (2022). Spatial proteogenomics reveals distinct and evolutionarily conserved hepatic macrophage niches. *Cell*, *185*(2), 379-39

6 e338. https://doi.org/10.1016/j.cell.2021.12.018
Hafemeister, C., & Satija, R. (2019). Normalization and variance stabilization of s ingle-cell RNA-seq data using regularized negative binomial regression. *Genome Biol*, *20*(1), 296. https://doi.org/10.1186/s13059-019-1874-1
Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch eff ects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nat Biotechnol*, *36*(5), 421-427. https://doi.org/10.1038/nbt.4091
Halpern, K. B., Shenhav, R., Matcovitch-Natan, O., Toth, B., Lemze, D., Golan, M., Massasa, E. E., Baydatch, S., Landen, S., Moor, A. E., Brandis, A., Giladi, A., Avihail, A. S., David, E., Amit, I., & Itzkovitz, S. (2017). Single-cell spatial r econstruction reveals global division of labour in the mammalian liver. *Nature*, *54 2*(7641), 352-356. https://doi.org/10.1038/nature21065
He, L., Pu, W., Liu, X., Zhang, Z., Han, M., Li, Y., Huang, X., Han, X., Li, Y., Liu, K., Shi, M., Lai, L., Sun, R., Wang, Q. D., Ji, Y., Tchorz, J. S., & Zhou, B. (2021). Proliferation tracing reveals regional hepatocyte generation in liver homeo stasis and repair. *Science*, *371*(6532). https://doi.org/10.1126/science.abc4346
Hickey, J. W., Becker, W. R., Nevins, S. A., Horning, A., Perez, A. E., Zhu, C., Zhu, B., Wei, B., Chiu, R., Chen, D. C., Cotter, D. L., Esplin, E. D., Weimer, A. K., Caraccio, C., Venkataraaman, V., Schurch, C. M., Black, S., Brbic, M., Cao, K., . . . Snyder, M. (2023). Organization of the human intestine at single-cell reso lution. *Nature*, *619*(7970), 572-584. https://doi.org/10.1038/s41586-023-05915-x
Hie, B., Bryson, B., & Berger, B. (2019). Efficient integration of heterogeneous si ngle-cell transcriptomes using Scanorama. *Nat Biotechnol*, *37*(6), 685-691. https://doi.org/10.1038/s41587-019-0113-3
Hu, J., Li, X., Coleman, K., Schroeder, A., Ma, N., Irwin, D. J., Lee, E. B., Shin ohara, R. T., & Li, M. (2021). SpaGCN: Integrating gene expression, spatial locat ion and histology to identify spatial domains and spatially variable genes by grap h convolutional network. *Nat Methods*, *18*(11), 1342-1351. https://doi.org/10.1038/s41592-021-01255-8
Hu, Y., Li, Y., Xie, M., Rao, M., Liu, Y. H., & Zhou, X. M. (2024). MaskGraphe ne: Advancing joint embedding, clustering, and batch correction for spatial transcr iptomics using graph-based self-supervised learning. *bioRxiv*. https://doi.org/10.1101/2024.02.21.581387
Isaacs, F. J., Hasty, J., Cantor, C. R., & Collins, J. J. (2003). Prediction and meas urement of an autoregulatory genetic module. *Proc Natl Acad Sci U S A*, *100*(1 3), 7714-7719. https://doi.org/10.1073/pnas.1332628100
Kingma, D. P., & Welling, M. (2013). Auto-Encoding Variational Bayes. https://doi.org/10.48550/arXiv.1312.6114
Kipf, T. N., & Welling, M. (2017). Semi-Supervised Classification with Graph Co nvolutional Networks. https://doi.org/10.48550/arXiv.1609.02907
Kleshchevnikov, V., Shmatko, A., Dann, E., Aivazidis, A., King, H. W., Li, T., El mentaite, R., Lomakin, A., Kedlian, V., Gayoso, A., Jain, M. S., Park, J. S., Ram ona, L., Tuck, E., Arutyunyan, A., Vento-Tormo, R., Gerstung, M., James, L., Ste gle, O., & Bayraktar, O. A. (2022). Cell2location maps fine-grained cell types in

spatial transcriptomics. *Nat Biotechnol*, *40*(5), 661-671. https://doi.org/10.1038/s41587-021-01139-4

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. R., & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat Methods*, *16*(12), 1289-1296. https://doi.org/10.1038/s41592-019-0619-0

Lake, B. B., Menon, R., Winfree, S., Hu, Q., Melo Ferreira, R., Kalhor, K., Barwinska, D., Otto, E. A., Ferkowicz, M., Diep, D., Plongthongkum, N., Knoten, A., Urata, S., Mariani, L. H., Naik, A. S., Eddy, S., Zhang, B., Wu, Y., Salamon, D., . . . Jain, S. (2023). An atlas of healthy and injured cell states and niches in the human kidney. *Nature*, *619*(7970), 585-594. https://doi.org/10.1038/s41586-023-05769-3

Lause, J., Berens, P., & Kobak, D. (2021). Analytic Pearson residuals for normalization of single-cell RNA-seq UMI data. *Genome Biol*, *22*(1), 258. https://doi.org/10.1186/s13059-021-02451-7

Lei Ba, J., Kiros, J. R., Hinton, G. E., Lei Ba, J., Kiros, J. R., & Hinton, G. E. (2016). Layer Normalization. *arXiv e-prints*. https://doi.org/10.48550/arXiv.1607.06450

Lemoine, G. G., Scott-Boyer, M. P., Ambroise, B., Perin, O., & Droit, A. (2021). GWENA: gene co-expression networks analysis and extended modules characterization in a single Bioconductor package. *BMC Bioinformatics*, *22*(1), 267. https://doi.org/10.1186/s12859-021-04179-4

Li, J., Chen, S., Pan, X., Yuan, Y., & Shen, H. B. (2022). Cell clustering for spatial transcriptomics data with graph neural networks. *Nat Comput Sci*, *2*(6), 399-408. https://doi.org/10.1038/s43588-022-00266-5

Li, Z., & Zhou, X. (2022). BASS: multi-scale and multi-sample analysis enables accurate cell type clustering and spatial domain detection in spatial transcriptomic studies. *Genome Biol*, *23*(1), 168. https://doi.org/10.1186/s13059-022-02734-7

Liu, W., Liao, X., Luo, Z., Yang, Y., Lau, M. C., Jiao, Y., Shi, X., Zhai, W., Ji, H., Yeong, J., & Liu, J. (2023). Probabilistic embedding, clustering, and alignment for integrating spatial transcriptomics data with PRECAST. *Nat Commun*, *14*(1), 296. https://doi.org/10.1038/s41467-023-35947-w

Long, Y., Ang, K. S., Li, M., Chong, K. L. K., Sethi, R., Zhong, C., Xu, H., Ong, Z., Sachaphibulkij, K., Chen, A., Zeng, L., Fu, H., Wu, M., Lim, L. H. K., Liu, L., & Chen, J. (2023). Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat Commun*, *14*(1), 1155. https://doi.org/10.1038/s41467-023-36796-3

Lopez, R., Li, B., Keren-Shaul, H., Boyeau, P., Kedmi, M., Pilzer, D., Jelinski, A., Yofe, I., David, E., Wagner, A., Ergen, C., Addadi, Y., Golani, O., Ronchese, F., Jordan, M. I., Amit, I., & Yosef, N. (2022). DestVI identifies continuums of cell types in spatial transcriptomics data. *Nat Biotechnol*, *40*(9), 1360-1369. https://doi.org/10.1038/s41587-022-01272-8

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., & Yosef, N. (2018). Deep generative modeling for single-cell transcriptomics. *Nat Methods*, *15*(12), 1053-1058. ht

tps://doi.org/10.1038/s41592-018-0229-2
Luecken, M. D., Buttner, M., Chaichoompu, K., Danese, A., Interlandi, M., Muell er, M. F., Strobl, D. C., Zappia, L., Dugas, M., Colome-Tatche, M., & Theis, F. J. (2022). Benchmarking atlas-level data integration in single-cell genomics. *Nat Methods*, *19*(1), 41-50. https://doi.org/10.1038/s41592-021-01336-8
Ma, Y., & Zhou, X. (2022). Spatially informed cell-type deconvolution for spatial transcriptomics. *Nat Biotechnol*, *40*(9), 1349-1359. https://doi.org/10.1038/s41587-022-01273-7
Maynard, K. R., Collado-Torres, L., Weber, L. M., Uytingco, C., Barry, B. K., Wi lliams, S. R., Catallini, J. L., 2nd, Tran, M. N., Besich, Z., Tippani, M., Chew, J., Yin, Y., Kleinman, J. E., Hyde, T. M., Rao, N., Hicks, S. C., Martinowich, K., & Jaffe, A. E. (2021). Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci*, *24*(3), 425-436. https://doi.org/10.1038/s41593-020-00787-0
Miller, B. F., Bambah-Mukku, D., Dulac, C., Zhuang, X., & Fan, J. (2021). Char acterizing spatial gene expression heterogeneity in spatially resolved single-cell tra nscriptomic data with nonuniform cellular densities. *Genome Res*, *31*(10), 1843-18 55. https://doi.org/10.1101/gr.271288.120
Moffitt, J. R., Bambah-Mukku, D., Eichhorn, S. W., Vaughn, E., Shekhar, K., Per ez, J. D., Rubinstein, N. D., Hao, J., Regev, A., Dulac, C., & Zhuang, X. (2018). Molecular, spatial, and functional single-cell profiling of the hypothalamic preopti c region. *Science*, *362*(6416), eaau5324. https://doi.org/10.1126/science.aau5324
Moffitt, J. R., Hao, J., Wang, G., Chen, K. H., Babcock, H. P., & Zhuang, X. (20 16). High-throughput single-cell gene-expression profiling with multiplexed error-r obust fluorescence in situ hybridization. *Proc Natl Acad Sci U S A*, *113*(39), 1104 6-11051. https://doi.org/10.1073/pnas.1612826113
Nagendran, M., Sapida, J., Arthur, J., Yin, Y., Tuncer, S. D., Anaparthy, N., Gupt a, A., Serra, M., Patterson, D., & Tentori, A. (2023). 1457 Visium HD enables sp atially resolved, single-cell scale resolution mapping of FFPE human breast cance r tissue. *Journal for ImmunoTherapy of Cancer*, *11*(Suppl 1). https://doi.org/10.1136/jitc-2023-SITC2023.1457
NT, H., Maehara, T., NT, H., & Maehara, T. (2019). Revisiting Graph Neural Net works: All We Have is Low-Pass Filters. *arXiv e-prints*. https://doi.org/10.48550/arXiv.1905.09550
Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of m odule detection methods for gene expression data. *Nat Commun*, *9*(1), 1090. https://doi.org/10.1038/s41467-018-03424-4
Satija, R., Farrell, J. A., Gennert, D., Schier, A. F., & Regev, A. (2015). Spatial r econstruction of single-cell gene expression data. *Nat Biotechnol*, *33*(5), 495-502. https://doi.org/10.1038/nbt.3192
Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., & Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. *Nat Genet*, *34*(2), 166-176. https://doi.org/10.1038/ng1165

Shang, L., & Zhou, X. (2022). Spatially aware dimension reduction for spatial tra nscriptomics. *Nat Commun*, *13*(1), 7203. https://doi.org/10.1038/s41467-022-34879-1

Stahl, P. L., Salmen, F., Vickovic, S., Lundmark, A., Navarro, J. F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J. O., Huss, M., Mollbrink, A., Linnarsso n, S., Codeluppi, S., Borg, A., Ponten, F., Costea, P. I., Sahlen, P., Mulder, J., Be rgmann, O., . . . Frisen, J. (2016). Visualization and analysis of gene expression i n tissue sections by spatial transcriptomics. *Science*, *353*(6294), 78-82. https://doi.org/10.1126/science.aaf2403

Stickels, R. R., Murray, E., Kumar, P., Li, J., Marshall, J. L., Di Bella, D. J., Arl otta, P., Macosko, E. Z., & Chen, F. (2021). Highly sensitive spatial transcriptomi cs at near-cellular resolution with Slide-seqV2. *Nat Biotechnol*, *39*(3), 313-319. https://doi.org/10.1038/s41587-020-0739-1

Stuart, T., & Satija, R. (2019). Integrative single-cell analysis. *Nat Rev Genet*, *20* (5), 257-272. https://doi.org/10.1038/s41576-019-0093-7

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillett e, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interp reting genome-wide expression profiles. *Proc Natl Acad Sci U S A*, *102*(43), 1554 5-15550. https://doi.org/10.1073/pnas.0506580102

Teng, H., Yuan, Y., & Bar-Joseph, Z. (2022). Clustering spatial transcriptomics da ta. *Bioinformatics*, *38*(4), 997-1004. https://doi.org/10.1093/bioinformatics/btab704

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kai ser, L., & Polosukhin, I. (2023). Attention Is All You Need. https://doi.org/10.48550/arXiv.1706.03762

[1]. Wang, G., Zhao, J., Yan, Y., Wang, Y., Wu, A. R., & Yang, C. (2023). Const ruction of a 3D whole organism spatial atlas by joint modelling of multiple sl ices with deep neural networks. *Nature Machine Intelligence*, *5*(11), 1200-121 3. https://doi.org/10.1038/s42256-023-00734-1

Wang, X., Allen, W. E., Wright, M. A., Sylwestrak, E. L., Samusik, N., Vesuna, S., Evans, K., Liu, C., Ramakrishnan, C., Liu, J., Nolan, G. P., Bava, F. A., & D eisseroth, K. (2018). Three-dimensional intact-tissue sequencing of single-cell tran scriptional states. *Science*, *361*(6400), eaat5691. https://doi.org/10.1126/science.aat5691

Wang, X., Dalkic, E., Wu, M., & Chan, C. (2008). Gene module level analysis: i dentification to networks and dynamics. *Curr Opin Biotechnol*, *19*(5), 482-491. https://doi.org/10.1016/j.copbio.2008.07.011

Wei, Y., Wang, Y. G., Jia, Y., Li, L., Yoon, J., Zhang, S., Wang, Z., Zhang, Y., Z hu, M., Sharma, T., Lin, Y. H., Hsieh, M. H., Albrecht, J. H., Le, P. T., Rosen, C. J., Wang, T., & Zhu, H. (2021). Liver homeostasis is maintained by midlobula r zone 2 hepatocytes. *Science*, *371*(6532). https://doi.org/10.1126/science.abb1625

Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell g ene expression data analysis. *Genome Biol*, *19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0

Xu, C., Lopez, R., Mehlman, E., Regier, J., Jordan, M. I., & Yosef, N. (2021). Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models. *Mol Syst Biol*, *17*(1), e9620. https://doi.org/10.15252/msb.20209620

Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., & Yao, J. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, *4*(10), 852-866. https://doi.org/10.1038/s42256-022-00534-z

Yang, Y., Cui, Y., Zeng, X., Zhang, Y., Loza, M., Park, S.-J., & Nakai, K. (2024). STAIG: Spatial Transcriptomics Analysis via Image-Aided Graph Contrastive Learning for Domain Exploration and Alignment-Free Integration. *bioRxiv*. https://doi.org/10.1101/2023.12.18.572279

Yu, J., Li, J., Yu, Z., & Huang, Q. (2020). Multimodal Transformer With Multi-View Visual Representation for Image Captioning. *IEEE Transactions on Circuits and Systems for Video Technology*, *30*(12), 4467-4480. https://doi.org/10.1109/tcsvt.2019.2947482

Yu, N., Zhang, D., Zhang, W., Liu, Z., Qiao, X., Wang, C., Zhao, M., Chao, B., Li, W., De Marinis, Y., & Gao, R. (2023). stGCL: A versatile cross-modality fusion method based on multi-modal graph contrastive learning for spatial transcriptomics. *bioRxiv*. https://doi.org/10.1101/2023.12.10.571025

Yuan, Z. (2024). MENDER: fast and scalable tissue structure identification in spatial omics data. *Nat Commun*, *15*(1), 207. https://doi.org/10.1038/s41467-023-44367-9

Zeng, H., Zhou, H., Srivastava, A., Kannan, R., Prasanna, V., Zeng, H., Zhou, H., Srivastava, A., Kannan, R., & Prasanna, V. (2019). GraphSAINT: Graph Sampling Based Inductive Learning Method. *arXiv e-prints*. https://doi.org/10.48550/arXiv.1907.04931

Zhang, B., He, P., Lawrence, J. E. G., Wang, S., Tuck, E., Williams, B. A., Roberts, K., Kleshchevnikov, V., Mamanova, L., Bolt, L., Polanski, K., Li, T., Elmentaite, R., Fasouli, E. S., Prete, M., He, X., Yayon, N., Fu, Y., Yang, H., . . . Teichmann, S. A. (2023). A human embryonic limb cell atlas resolved in space and time. *Nature*. https://doi.org/10.1038/s41586-023-06806-x

Zhang, D., Yu, N., Li, W., Sun, X., Zou, Q., Li, X., Liu, Z., Yuan, Z., Zhang, W., & Gao, R. (2024). stMMR: accurate and robust spatial domain identification from spatially resolved transcriptomics with multi-modal feature representation. *bioRxiv*. https://doi.org/10.1101/2024.02.22.581503

Zhao, E., Stone, M. R., Ren, X., Guenthoer, J., Smythe, K. S., Pulliam, T., Williams, S. R., Uytingco, C. R., Taylor, S. E. B., Nghiem, P., Bielas, J. H., & Gottardo, R. (2021). Spatial transcriptomics at subspot resolution with BayesSpace. *Nat Biotechnol*, *39*(11), 1375-1384. https://doi.org/10.1038/s41587-021-00935-2

Zheng, G. X., Terry, J. M., Belgrader, P., Ryvkin, P., Bent, Z. W., Wilson, R., Ziraldo, S. B., Wheeler, T. D., McDermott, G. P., Zhu, J., Gregory, M. T., Shuga, J., Montesclaros, L., Underwood, J. G., Masquelier, D. A., Nishimura, S. Y., Schnall-Levin, M., Wyatt, P. W., Hindson, C. M., . . . Bielas, J. H. (2017). Massively p

arallel digital transcriptional profiling of single cells. *Nat Commun*, *8*(1), 14049. https://doi.org/10.1038/ncomms14049
Zhou, X., Dong, K., & Zhang, S. (2023). Integrating spatial transcriptomics data across different conditions, technologies and developmental stages. *Nat Comput Sci*, *3*(10), 894-906. https://doi.org/10.1038/s43588-023-00528-w
Zhu, J., Sun, S., & Zhou, X. (2021). SPARK-X: non-parametric modeling enables scalable and robust detection of spatial expression patterns for large spatial transcriptomic studies. *Genome Biol*, *22*(1), 184. https://doi.org/10.1186/s13059-021-02404-0

Methods

The STEP backbone model (BBM)

For a batch of gene expression $\boldsymbol{X} \in \mathbf{R}_{+}^{N\times G}$ with N observations and G sequenced genes, the n-th gene expression data $\mathbf{x}_n \in \boldsymbol{X}$ is firstly performed by a layer normalization, and then it is mapped to a length K sequence corresponding to the gene module $\boldsymbol{M}_n \in \mathbf{R}^{K\times d}$ by a series of randomly initialized linear transformations $\mathbf{W} \in \mathbf{R}^{G\times d\times K}$, where $d \ll G$ and each transformation aims to learn one gene module. And $\mathbf{x}_n$ is followingly input to a ReLU function

$$\boldsymbol{M}_n^T = \mathrm{ReLU}(\mathrm{LayerNorm}(\mathbf{x}_n)\mathbf{W}) \tag{1}$$

Then we concatenate a cls-token $\mathbf{z}^{cls} \in \mathbf{R}^D$ to $\boldsymbol{M}_n$ to extract the biological information behind $\mathbf{x}_n$ and refer to the final low-dimensional representation of $\mathbf{x}_n$. Denoting $\boldsymbol{M}'_n = \{\mathbf{z}^{cls}, \boldsymbol{M}_n\}$, we up-sample (expand) $\boldsymbol{M}'_n$ to $\boldsymbol{M}_n^{expd}$ with dimension D by another affine transformation. The purpose of up-sampling is to counteract any potential negative regulation within each gene module. Given that the ReLU function only outputs positive values, every element within each gene module is positive. Also, the attention scores are exclusively positive due to the properties of the softmax function. Next, we feed $\boldsymbol{M}_n^{expd}$ with an additive positional embedding $E_{pos} \in R^{(1+K)\times d}$ to L Transformer encoders $f_{enc}^{(l)}(\cdot)$, $l = 1, 2, \ldots, L$ (default $L = 3$)

$$\boldsymbol{M}_n^{expd} = \boldsymbol{M}'_n \boldsymbol{W}_{expd} + \mathbf{b}_{expd}, \quad \boldsymbol{W}_{expd} \in \mathbf{R}^{d\times D}, \mathbf{b}_{expd} \in \mathbf{R}^D \tag{2}$$

$$\boldsymbol{O}_n = \left(\prod_{l=1}^{L} f_{enc}^{(l)}\right)\left(\boldsymbol{M}_n^{expd} + \boldsymbol{E}_{pos}\right) \tag{3}$$

Here, the positional embedding $\boldsymbol{E}_{pos}$ is used to differentiate each gene modules. Each Transformer encoder $f_{enc}^{(l)}$ consists of a multi-head self attention module and a feed forward network, the multi-head attention part can be described formally as

$$\{\mathbf{Q}_n, \mathbf{K}_n, \mathbf{V}_n\} = \boldsymbol{M}_n^{epd}\{\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V}\}, \quad \mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V} \in \mathbf{R}^{D\times D\times n_h} \tag{4}$$

$$\mathbf{A}_n = \text{softmax}\left(\frac{\mathbf{Q}_n^T\mathbf{K}_n}{\sqrt{D}}\right) \tag{5}$$

$$\boldsymbol{O}_n^{attn} = \left(||_{i=1}^{n_h}\left(\mathbf{A}_n(i)\mathbf{V}_n(i)\right)\right)\boldsymbol{W}_O, \boldsymbol{W}_O \in \mathbf{R}^{Dn_h\times D} \tag{6}$$

where n_h is the number of head in multi-head self-attention, $\text{softmax}(\mathbf{x}) = \frac{e^{-\mathbf{x}}}{\sum_{i=1}^{n} e^{-x_i}}, \mathbf{x} \in \mathbf{R}^d$ and $\mathbf{A}_n \in \mathbf{R}^{K\times K\times n_h}$ refers to the attention score matrix for each attention head. The product of $\mathbf{A}_n(i)$ and $\mathbf{V}_n(i)$ are then concatenated in head-wise and multiplied by another weight $\boldsymbol{W_O}$ to obtain the final output $\boldsymbol{O}_n^{attn} \in \mathbf{R}^{(K+1)\times D}$ of a multi-head attention module. And the feedforward network consists of two fully connected layers and a non-linear activation function

$$\begin{aligned}\boldsymbol{O}_n &= \text{Feedforward}(\boldsymbol{O}_n^{attn}) \\ &= \text{ReLU}(\boldsymbol{O}_n^{attn}\boldsymbol{W}_1 + \mathbf{b}_1)\boldsymbol{W}_2 + \mathbf{b}_2\end{aligned} \tag{7}$$

where $\boldsymbol{W}_1 \in \mathbf{R}^{D\times 4D}$, $\boldsymbol{W}_2 \in \mathbf{R}^{4D\times D}$, and we extract the first column of $\boldsymbol{O}_n$ corresponding to the previous cls-token $\mathbf{z}^{cls}$ as the non-standardized embedding, denoted as $\mathbf{h}_n$.

Afterwards, based on the widely used reparameterization trick in all VAE models, the $\mathbf{h}_n$ is processed by a Readout module $f_{rout}(\cdot)$ to estimate the mean $\boldsymbol{\mu}_n$ and the logarithm of variance $2\log\boldsymbol{\sigma_n}$ of the standardized low-dimensional embedding $\mathbf{z}_n \sim N(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n^2)$.

$$\begin{aligned}\text{Logits}_n &= f_{rout}(\mathbf{h}_n) \\ &= \text{LayerNorm}\Big(\text{Dropout}\big(\text{GELU}(\mathbf{h}_n\boldsymbol{W}_{rout} \\ &\quad + \mathbf{b}_{rout})\big)\Big)\end{aligned} \tag{8}$$

$$\boldsymbol{\mu}_n = \text{logits}_n\boldsymbol{W}_{\boldsymbol{\mu}} + \mathbf{b}_{\boldsymbol{\mu}} \tag{9}$$

$$\log\boldsymbol{\sigma}_n = \text{logits}_n\boldsymbol{W}_{\boldsymbol{\sigma}} + \mathbf{b}_{\boldsymbol{\sigma}} \tag{10}$$

where $\boldsymbol{W}_{rout} \in \mathbf{R}^{D\times D}, \boldsymbol{W}_{\boldsymbol{\mu}} \in \mathbf{R}^{D\times d}$. $\mathbf{z}_n$ is then put to a non-linear decoder $f_{dec}(\cdot)$ to obtain the parameter set of the dispersion, which is shared across cells/spots, and the zero-inflation $(\mathbf{r}, \boldsymbol{\pi}_n)$ of the ZINB distribution, as well as the mean expression $\boldsymbol{\mu}_n^{expr} \in \mathbf{R}^G$,

$$\begin{aligned}\text{logits}_n^{dec} &= f_{dec}(\mathbf{z}_n) \\ &= \text{ReLU}\big(\text{BatchNorm}(\mathbf{z}_n\boldsymbol{W}_{dec} + \mathbf{b}_{dec})\big)\end{aligned} \tag{11}$$

$$\boldsymbol{\mu}_n^{expr} = \text{library}(\mathbf{x}_n) \cdot \text{softmax}\left(\text{logits}_n^{dec} \boldsymbol{W}_{\log \boldsymbol{\mu}^{expr}} + \mathbf{b}_{\log \boldsymbol{\mu}^{expr}}\right) \tag{12}$$

$$\text{logit}(\boldsymbol{\pi}_n) = \text{logits}_n^{dec} \boldsymbol{W}_{\boldsymbol{\pi}} + \mathbf{b}_{\boldsymbol{\pi}}, \tag{13}$$

where $\{\boldsymbol{W}_{dec}^T, \mathbf{b}_{dec}^T\} \in \mathbf{R}^{D\times(d+1)}$, $\{\boldsymbol{W}_{\log \boldsymbol{\mu}^{expr}}^T, \mathbf{b}_{\log \boldsymbol{\mu}^{expr}}^T\}$ and $\{\boldsymbol{W}_{\boldsymbol{\pi}}^T, \mathbf{b}_{\boldsymbol{\pi}^T}\} \in \mathbf{R}^{G\times(D+1)}$, and $\text{logit}(x) = \log\left(\frac{x}{1-x}\right)$, $\text{library}(\mathbf{x}) = \sum_{i=1}^{m} x_i$ when $\mathbf{x} \in \mathbf{R}^m$. The dispersion $\mathbf{r}$ is randomly initialized and learned during the training. After training, we can sample the reconstruction/imputation $\hat{\mathbf{x}}_n$ for the gene expression data $\mathbf{x}_n$ from $\text{ZINB}(\boldsymbol{\mu}_n, \mathbf{r}, \boldsymbol{\pi}_n)$. We denote the probability density function (p.d.f.) of $\mathbf{z}_n \sim N(\boldsymbol{\mu}_n, \boldsymbol{\sigma}_n^2)$ as $q_\phi(\mathbf{z}_n)$, where ϕ refers to the parameters to learn, and let the p.d.f. of the standard normal distribution, with the dimension equal to that of $\mathbf{z}_n$, be $p(\mathbf{z}_n)$. We train the backbone model by minimizing the objective function:

$$\mathcal{L}(\boldsymbol{X}) = -\frac{1}{N}\sum_{n=1}^{N}\left[\log p_{\hat{\mathbf{x}}_n \sim \text{ZINB}(\boldsymbol{\mu}_n^{expr}, \mathbf{r}, \boldsymbol{\pi}_n)}(\mathbf{x}_n)\right] + \beta \text{KL}\left(q_\phi(\mathbf{z}_n) || p(\mathbf{z}_n)\right) \tag{14}$$

The backbone model is then extended with another two models to perform more complex analysis and the combined model is trained by different objectives for different tasks.

Batch effect model (BEM)

We propose a novel and simple batch-wise layer normalization named BatchAware LayerNorm and use it to replace the LayerNorm before gene module transformations (in Eq.1) when it comes to multi-batches data analysis. We firstly use the one-hot encoding of each batch indicator $b \in B$ and randomly initialize a batch-embedding $\boldsymbol{E}_B \in \mathbf{R}^{|B|\times D}$, so that our model can be guided to produce a batch-invariant low-dimensional representation $\mathbf{z}_n$. The computation of BatchAwareLayerNorm can be described as:

$$\begin{aligned} \mathbf{x}_n' &= \text{BatchAwareLayerNorm}(\mathbf{x}_n, b, \boldsymbol{E}_B) \\ &= \text{onehot}(b)\boldsymbol{E}_B \boldsymbol{W}_{\boldsymbol{\mu}_b} + \text{onehot}(b)\boldsymbol{E}_B \mathbf{b}_{\boldsymbol{\mu}_b} \\ &+ \left(\text{onehot}(b)\boldsymbol{E}_B \boldsymbol{W}_{\boldsymbol{\sigma}_b} + \text{onehot}(b)\boldsymbol{E}_B \mathbf{b}_{\boldsymbol{\sigma}_b}\right) \\ &\odot \frac{\mathbf{x}_n - \mu(\mathbf{x}_n)}{\sigma(\mathbf{x}_n)} \end{aligned} \tag{15}$$

where $\{\boldsymbol{W}_{\boldsymbol{\mu}_b}^T, \mathbf{b}_{\boldsymbol{\mu}_b}^T\}, \{\boldsymbol{W}_{\boldsymbol{\sigma}_b}^T, \mathbf{b}_{\boldsymbol{\sigma}_b}^T\} \in \mathbf{R}^{G\times(D+1)}$. To keep the consistency in the decoding part, we perform BatchAwareScale on $\mathbf{z}_n$ to retrieve the batch variations before it is sent to decoder $f_{dec}(\cdot)$:

$$\begin{aligned} \boldsymbol{z}_n' &= \text{BatchAwareScale}(\mathbf{z}_n, b, \boldsymbol{E}_b) \\ &= \text{GELU}((\mathbf{z}_n || \text{onehot}(b)\boldsymbol{E}_B)\boldsymbol{W}_3 + \mathbf{b}_3)\boldsymbol{W}_4 + \mathbf{b}_4, \end{aligned} \tag{16}$$

where $\{\boldsymbol{W}_3^T, \mathbf{b}_3^T\} \in \mathbf{R}^{4D\times(D+1)}, \{\boldsymbol{W}_4^T, \mathbf{b}_4^T\} \in \mathbf{R}^{D\times(4D+1)}$. These two paired operations consist of the BEM. The combination of BBM and BEM is still trained with the basic objective (Eq.14).

If the cell-type annotations set C can be accessed, we can use them as labels to train a cell-type classifier $f_{clsf}(\cdot)$ to perform a much more precise integration and produce the centroids for all cell types. Specifically, after the model is trained without annotation information, we randomly initialize the set of centoids $\boldsymbol{Z}^{anchor} \in \mathbf{R}^{|C|\times d}$ and add the cell-type classifier $f_{clsf}(\cdot)$ after reading out the module $f_{rout}(\cdot)$. For a mini-batch of low-dimensional representations $\boldsymbol{Z}_{\mathcal{B}} \subseteq \boldsymbol{Z} = \{\mathbf{z}_n\}_{n=1}^N$, the predicted probability matrices of cell types $\widehat{\boldsymbol{C}}_{\mathcal{B}} \in \mathbf{R}^{|\mathcal{B}|\times|C|}$ and $\widehat{\boldsymbol{C}}^{anchor} \in \mathbf{R}^{|C|\times|C|}$ for mini-batches data and $\boldsymbol{Z}^{anchor}$, respectively, would be inferred together as follows,

$$\begin{aligned} \widehat{\boldsymbol{C}}_{\mathcal{B}} || \widehat{\boldsymbol{C}}^{anchor} &= f_{clsf}(\boldsymbol{Z}_{\mathcal{B}} || \boldsymbol{Z}^{anchor}) \\ &= \text{softmax}(\text{LayerNorm}\Big((\boldsymbol{Z}_{\mathcal{B}} || \boldsymbol{Z}^{anchor})\boldsymbol{W}_5 + \mathbf{b}_5)\boldsymbol{W}_6 + \mathbf{b}_6\Big), \end{aligned} \tag{17}$$

where $\{\boldsymbol{W}_5^T, \mathbf{b}_5^T\} \in \mathbf{R}^{D\times(D+1)}, \{\boldsymbol{W}_6^T, \mathbf{b}_6^T\} \in \mathbf{R}^{|C|\times(D+1)}$.

For training the cell-type classifier meanwhile refining the embedding of the BBM, we set the learning rate of the modules except f_{clsf} to a very low level (10^{-5}) and minimize another objective $\mathcal{L}_{clsf}$:

$$\begin{aligned} \mathcal{L}_{clsf}(\boldsymbol{X}, B, C) = \frac{1}{N}\sum_{n=1}^{N} \text{CE}\big(\hat{\mathbf{c}}_n || \widehat{\boldsymbol{C}}^{anchor}, \mathbf{c}_n || \boldsymbol{C}^{anchor}\big) \\ + \beta \text{KL}\Big(q_\phi(\mathbf{z}_n) || p(\mathbf{z}_n)\Big), \end{aligned} \tag{18}$$

where CE represents cross-entropy, $\mathbf{c}_n$ and $\boldsymbol{C}^{anchor}$ represents the annotations of n-th cell and anchor points, respectively.

Spatial model (SpM)

To understanding spatial context in tissues, we propose using STEP to generate spatially informed embeddings with the help of SpM, denoted as $g(\cdot)$. STEP first constructs the spatial neighboring graphs for each tissue. We denote the SRT of one section as $\boldsymbol{Y}$. For 10x Visium data, we construct a spatial neighboring graph $\boldsymbol{A}_{\boldsymbol{Y}}$ using hexagonal grids aligned in each tissue. For single-cell resolution SRT data, we compute kNN graph(s) based on spatial locations in each tissue. The graphs have binary edges and include self-loops in $\boldsymbol{A}_{\boldsymbol{Y}}$. Next, we apply a vanilla GCN (without residual connection and activation) as a spatial low-pass filter for non-single-cell resolution SRT data, or a smooth operator like a parametric graph Laplacian for single-cell resolution SRT data. SpM filters or smooths the non-standardized embedding $\boldsymbol{H} = \{\mathbf{h}_s\}_{s=1}^S$ in (Eq.3) before

passing it to the Readout module $f_{rout}(\cdot)$. The output of SpM is denoted as $\widetilde{\boldsymbol{H}}$. SpM utilizes L (default=4) graph convolutional layers for computation,

$$\begin{aligned}\widetilde{\boldsymbol{H}} &= g(\boldsymbol{A_Y}, \{\mathbf{h}_s\}_{s=1}^{S}) \\ &= \text{LayerNorm}\left(\left(\boldsymbol{D_Y}^{-\frac{1}{2}}\boldsymbol{A_Y}\boldsymbol{D_Y}^{-\frac{1}{2}}\right)^{L}\boldsymbol{H}\prod_{l=1}^{L}\boldsymbol{W}_g^{(l)}\right),\end{aligned} \tag{19}$$

where $\boldsymbol{W}_g^{(l)} \in \mathbf{R}^{D\times D}$ is the weight of l-th graph convolutional layer, $l \in \{1, 2, \dots, L\}$.

As mentioned before, to ensure proper running time and direct access to spatial domains for non-single-cell resolution SRT data, we employ the "fast mode", an architecture similar to the SimCSE for end-to-end training with the incorporation of contrastive loss. The forward process in the BBM of unstandardized embedding is branched, with one path leading to the readout and decoder, and the other path being filtered/smoothed by SpM before joining the same path as the former. In this fast mode, STEP is trained by minimizing the losses generated by these two forward paths and the contrastive loss of the filtered path and the other path. The contrastive loss servers as a regularization term to control the strength of spatial filtering or smoothing. Formally, for each $\mathbf{y}_s \in Y$, denoting the non-filtered branch as $\mathbf{y}_s' \sim \text{ZINB}\left(\boldsymbol{\mu}_s^{expr}, \mathbf{r}, \boldsymbol{\pi}_s \middle| \boldsymbol{A_Y}\right)$ and filtered branch as $\mathbf{y}_s'' \sim \text{ZINB}\left(\boldsymbol{\mu}_s^{expr}, \mathbf{r}, \boldsymbol{\pi}_s \middle| \boldsymbol{A_Y}\right)$, we have the objective $\mathcal{L}_{st}$ as follow:

$$\begin{aligned}\mathcal{L}_{st}(\boldsymbol{Y}, \boldsymbol{A_Y}) = -\frac{1}{S}\sum_{s=1}^{S}\Big[&\log p_{\mathbf{y}_s' \sim \text{ZINB}(\boldsymbol{\mu}_s^{expr}, \mathbf{r}, \boldsymbol{\pi}_s)}(\mathbf{y}_s) \\ &+ \log p_{\mathbf{y}_s'' \sim \text{ZINB}(\boldsymbol{\mu}_s^{expr}, \mathbf{r}, \boldsymbol{\pi}_s | \mathbf{A_Y})}(\mathbf{y}_s)\Big] \\ &+\text{CE}\left(\frac{\boldsymbol{H}^T\widetilde{\boldsymbol{H}}}{T}, \mathbf{1}_{\boldsymbol{S}}\right) + \\ \beta\text{KL}\left(q_\phi(\mathbf{z}_s)||p(\mathbf{z}_s)\right),\end{aligned} \tag{20}$$

where T represents the temperature parameter; $\mathbf{1}_{\boldsymbol{S}} \in \boldsymbol{R}^S$ is a vector of which the elements are all 1. We also utilize the SAINT node sampler (Zeng et al., 2019) to further boost the training speed where the sampler randomly samples several nodes to form a sub-graph in each iteration.

Revealing targeted biological heterogeneities

STEP effectively balances all kinds of variations from the SRT and scRNA-seq data, and condenses them into the low-dimensional embedding: $\boldsymbol{H}$ (non-standardized), $\boldsymbol{Z}$ (standardized), and $\widetilde{\boldsymbol{Z}}$ (spatially smoothed and standardized). With these kinds of embeddings, the targeted biological heterogeneities, cell types (states) and spatial domains, can be easily revealed by commonly used community detection or clustering algorithms. This allows users to run STEP once to try multiple domain numbers, just like the standard scRNA-seq data analysis workflow. For accessing sub cell types and sub spatial domains, user can achieve them by simply increasing the number of clusters

to detect or run clustering method in a hierarchical style.

A special case is the integrative analysis for a pair of annotated scRNA-seq data and non-single-cell resolution SRT data. We propose a cell type deconvolution algorithm relying on their co-embedding. Particularly, in the deconvolution process of STEP, we denote scRNA-seq dataset as $\boldsymbol{X} \in \mathbf{R}_+^{N\times G}$ and SRT dataset as $\boldsymbol{Y} \in \mathbf{R}_+^{S\times G}$. The co-embedding result (using non-standardized embedding) is denoted as $H^T = \{(\boldsymbol{H}^{sc})^T; (\boldsymbol{H}^{st})^T\} \in R^{D\times(N+S)}$. Assuming the spatial domains in tissues are identified, our goal is to estimate the cell-type composition $\boldsymbol{P} \in \mathbf{R}_+^{S\times|C|}$ for the SRT dataset, where C represents the cell type annotation set. So, the first step is to generate cell-type anchors $\boldsymbol{H}^{anchor} \in \mathbf{R}^{|C|\times D}$ by averaging the non-standardized embedding $H^{sc} = \{\mathbf{h}_s^{sc}\}_{n=1}^N$ based on the cell types. Then the cell-type signatures $\boldsymbol{X}^{sig} \in \mathbf{R}_+^{|C|\times G}$ are computed by normalizing and averaging the gene expressions in reference scRNA-seq dataset to a probability matrix based on cell types. For each spot (s-th spot), STEP concatenates each $\mathbf{h}_s^{st} \in \mathbf{R}^D$ and cell-type-anchors $\boldsymbol{H}^{anchor}$, and feeds them to the new Transformer encoders f_{deconv}^{ct}. The output is split to $\boldsymbol{h}_s^{st} \in \mathrm{R}^D$ and $\boldsymbol{\mathcal{H}}_s^{anchor} \in \mathbf{R}^{|C|\times D}$ where $\boldsymbol{\mathcal{H}}_s^{anchor}$ is considered as the spot-specific cell-type-anchors and it is constrained to be as similar with $\boldsymbol{\mathcal{H}}_s^{anchor}$ as possible. They are further fed as query and key into a sparse attention module f_{attn} to calculate the non-negative cell-type similarity matrix $\boldsymbol{P_H} = f_{attn}(\boldsymbol{\mathcal{H}}^{\mathrm{st}}, \boldsymbol{\mathcal{H}}_s^{anchor}) \in \mathbf{R}^{S\times|C|}$ in latent space where $\boldsymbol{\mathcal{H}}^{st} = \{\boldsymbol{h}_s^{st}\}_{s=1}^S$ and the "convolution" of scRNA-seq data embedding in latent space is represented as $\boldsymbol{P_H} f_{rout}(\boldsymbol{\mathcal{H}}_s^{anchor})$. This convoluted result is then sent to decoder to compute reconstruction of data. As its name indicates, each entry of $\boldsymbol{P_H}$ represents the similarity between some cell types and some spots.

The estimation $\widehat{\boldsymbol{P}}$ of cell-type composition $\boldsymbol{P}$ in gene-expression level is transformed from the similarity matrix $\boldsymbol{P_H}$: $\widehat{\boldsymbol{P}} = f_\theta(\boldsymbol{P_H})$ by some transformation f_θ. Since the embedding is informative enough and the subtle difference will be complemented by minimizing the reconstruction error term, we want the transformation f_θ won't change $\boldsymbol{P_H}$ too much, and also in our assumption, $\boldsymbol{P_H}$ is defined as cell-type similarity matrix so that it should recognize what cell types are included in the spot, so we model the transformation f_θ as a element-wise scale function: $f_\theta(\boldsymbol{P_H}) = \boldsymbol{\gamma}^{ct} \circ \boldsymbol{P_H}$ where $\boldsymbol{\gamma}^{ct}$ models the cell-type-wise scale.

Since the gene expression data is modeled by $\mathrm{ZINB}(\boldsymbol{\mu}, \mathbf{r}, \boldsymbol{\pi})$, where parameters $(\boldsymbol{\mu}, \mathbf{r}, \boldsymbol{\pi})$ refer to mean, dispersion and zero-inflation rate, respectively. The zero-inflation parameter is modeled as a cell-dependent or spot-dependent parameter. We additionally model a "convolution" in the gene expression level: $\sum_{c\in C} \widehat{\boldsymbol{P}}_s \boldsymbol{X}_c^{sig}$ ($\hat{P}_s$ is the s-th row of $\widehat{\boldsymbol{P}}$ and $\boldsymbol{X}_c^{sig}$ is the c-th column of $\boldsymbol{X}^{sig}$). Formally, for the s-th spot $\mathbf{y}_s$, the NB part is directly parameterized by the shared dispersion $\mathbf{r}$ and two mean modelling ways as $\mathrm{NB}\left(\sum_{c\in C} \widehat{\boldsymbol{P}}_s \boldsymbol{X}_c^{sig}, \mathbf{r}\right)$ and $\mathrm{NB}\left(f_{dec}\left(\boldsymbol{P_H} f_{rout}(\boldsymbol{\mathcal{H}}_s^{anchor})\right), \mathbf{r}\right)$, the zero-inflated rate $\boldsymbol{\pi}_s$ is inferred by the embedding $\mathbf{z}_s^{st}$ of this spot. As a result, the objective $\mathcal{L}_{deconv}$ of deconvolution process can be written as:

$$
\begin{aligned}
&\mathcal{L}_{deconv}(\boldsymbol{Y}, \boldsymbol{H}^{st}, \boldsymbol{H}^{anchor}) \\
&= -\frac{\alpha}{S}\sum_{s=1}^{S}\Big[\log p_{\mathbf{y}_s' \sim \mathrm{ZINB}\left(\sum_{c\in C}\widehat{\boldsymbol{P}}_s \boldsymbol{X}_c^{sig}, \mathbf{r}, \boldsymbol{\pi}_s\right)}(\mathbf{y}_s) \\
&+ \log p_{\mathbf{y}_s'' \sim \mathrm{ZINB}\left(f_{dec}^{st}\left(\boldsymbol{P}_{\boldsymbol{H}} f_{rout}(\boldsymbol{\mathcal{H}}_s^{anchor})\right), \mathbf{r}, \boldsymbol{\pi}_s\right)}(\mathbf{y}_s)\Big] \\
&+ \mathrm{CE}\left(\langle \boldsymbol{H}^{anchor}, \boldsymbol{\mathcal{H}}_s^{anchor}\rangle, \boldsymbol{I}_{|C|}\right) \\
&+ \frac{1}{|\mathcal{D}|}\sum_{d\in\mathcal{D}}\sum_{s=1}^{S}\left(\widehat{\boldsymbol{P}}_s - \overline{\widehat{\boldsymbol{P}}}_d\right)^2 \mathbb{1}[\mathcal{D}(s) = d],
\end{aligned}
\tag{21}
$$

where α is a hyper-parameter (default 0.1), $\mathcal{D}$ represents the set of identified spatial domains, $\mathcal{D}(s)$ means the spatial domain to which spot s belongs, $\boldsymbol{I}_{|C|} \in \mathbf{R}^{|C|\times|C|}$ represents the identity matrix, and $\overline{\widehat{\boldsymbol{P}}}_d$ represents the average estimated cell-type composition of spots in spatial domain d.

Hyper-parameter settings

We use Adam as our optimizer and set the learning rate at 0.001, and the validation rate is set at 0.2. The number of training epochs is set to 400 and Early Stopping is enabled in default. The number of gene modules K is set to 32 in default, and this value is tested in almost all dataset studied in this paper and shows great robustness. However, a high value of K means a longer running time, and this value can be decrease to 10-12 for some simple datasets (like Human DLPFC dataset) with only several cell types or spatial domains. Similarly, the dimensions of H and Z are set as 64 and 30, respectively, but they can be decreased to 40 and 20 when the loading and complexity of data are not so heavy, like Mouse medial prefrontal cortex data by STARmap. Overall, STEP is robust to the setting of these hype-parameters. The vital setting of hyper-parameters is about the value of β (default 10^{-2}), in the range from 10^{-2} to 10^{-4}, the performances of STEP is steady and well. But for the best performance in a certain dataset, like trying to reach the best batch-removal performance, needs to tune (often increase the β value from 10^{-2}) the β carefully since β with too high value may leads to overfit on the bottleneck and collapse of the model. On the contrary, STEP without variational can also work well and is also good at trajectory conservation but the points cannot gather too tight in the views of UMAP and suffer from the drop of performance in the batch correction score, which might lead to mis-cluster of cell types (spatial domains is quite robust to the value of β). In conclusion, too high value of β may lead to model collapse while no variational STEP may have trouble to cluster cells with the common clustering methods.

Focusing on SRT data, in the fast mode, the number iterations is set to 2000 for single-batch, and number of sampled spots/cells is set to 2048. The graph convolutional layers in spatial model are set to 6 in default. This hyper-parameter works for the

filtering or smoothing range in the tissue, so different values will influence the smoothness and continuity of identified domains. By our tests, performances of STEP are steady with 4 to 6 graph convolutional layers. An important model setting which cause overfit of the model is whether to use BatchNorm in decoder. There will be no performance drop for STEP without BatchNorm since our BatchAwareScale operations can do the similar job with BatchNorm in a categorical batch-wise rather than mini-batch-wise. But STEP can be trained much faster with much less epochs with BatchNorm. This phenomenon wildly acknowledged in the field of deep learning. However, when it comes to the occurrence of an extremely unbalanced counts levels, or sequencing depths, of the batches in a dataset, using BatchNorm tends to suffer from overfit. This is understandable, since BatchNorm is doing a mini-batch-wise normalization where the mini-batch in training STEP is several sampled sections, and such a universal normalization is quite incompatible with the huge difference of counts levels. Otherwise, when counts levels are balanced, meaning similar expression level, BatchNorm as a universal normalization can promote to the convergence of STEP.

Metrics

Cell-type level Multi-batch Integration Metrics

The whole evaluation process follows the scib (Luecken et al., 2022) pipeline with the same metrics. Evaluations in scib pipeline consist of two main aspects: metrics for the Bio conservation, including normalized mutual information (NMI), adjusted rand index (ARI), average silhouette width (ASW) measured by cell-type (ASW cell-type), isolated label scores (F1 and ASW) and graph local inverse Simpson index(LISI) for cell-type; and metrics for the Batch correction, including average silhouette width measured by batch (ASW batch), principal component regression (PCR), kBET and iLISI.

Except the STEP's results, all the other results are directly obtained from the original study. And all graph-based metrics (in this test) are evaluated on the kNN graph built on the outputted embedding of each method with same k values.

Normalized Mutual Information (NMI)

NMI evaluates the match level between clustering results and ground truth by examining their mutual information. NMI normalizes this based on their individual entropies, generally resulting in a value between 0 to 1. High NMI means a substantial mutual relationship is present, usually indicating a good clustering that matches the ground-truth labels. Formally, for a clustering result $\boldsymbol{X}$ and the ground truth $\boldsymbol{Y}$, NMI can be computed as

$$\mathrm{NMI}(\boldsymbol{X}, \boldsymbol{Y}) = \frac{2 * I(\boldsymbol{X}; \boldsymbol{Y})}{H(\boldsymbol{X}) + H(\boldsymbol{Y})}$$

where $I(\cdot;\cdot)$ represents the mutual information and $H(\cdot)$ represents the entropy.

Adjusted Rand Index (ARI)

ARI is a metric for evaluating the correspondence between clustering results and ground truth labels. It is normalized between -1 and 1. A high positive ARI signifies that the clustering result aligns well with the ground-truth labels, with 1 indicating the perfect match up and negative values indicating independent labels. Formally, for a set of n elements and two partitions of these elements $\boldsymbol{a}$ and $\boldsymbol{b}$. ARI can be computed as

$$\mathrm{ARI}(\boldsymbol{a},\boldsymbol{b}) = \frac{\binom{n}{2}\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}{\frac{1}{2}\left[\sum_i\binom{a_i}{2}+\sum_j\binom{b_j}{2}\right] - \left[\sum_i\binom{a_i}{2}\sum_j\binom{b_j}{2}\right]/\binom{n}{2}}$$

where a_i and b_j are the sum of pairs in the same set within $\boldsymbol{a}$ and $\boldsymbol{b}$.

Silhouette Coefficient

The Silhouette Coefficient is a metric used to appraise the consistency within clusters in a data set. For each sample in the data set, the Silhouette Coefficient is evaluated using the mean intra-cluster distance (average distance of the sample to all other data points in its assigned cluster) and the mean nearest-cluster distance (average distance of the sample to all samples of the next nearest cluster). It's defined for each data i and is composed of intra-cluster distance $a(i)$ and mean nearest-cluster distance $b(i)$:

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}$$

LISI

LISI is a measure utilized to evaluate the quality of clustering results in mixed samples. It inspects the diversity of labels in the local neighborhood of each sample, enabling an assessment of the mix of batches or cell types in a reduced space. For each data i, C_i denotes the cluster or the category that the sample belongs to, and $N(i)$ represents the data that are neighboring to i. The LISI score is individually calculated for each sample i through the inverse Simpson concentration of its neighboring samples, referred to as $\mathrm{LISI}(i)$. The overall Local Inverse Simpson Index for the entire dataset is determined by taking the harmonic mean of the $\mathrm{LISI}(i)$ values across all samples in the dataset. Achieving a higher LISI value is indicative of better mixing of categories within the dataset,

$$\mathrm{LISI}(i) = \frac{1}{\sum_{j\in C}(p_{ij}^2)}$$

where p_{ij} denotes the proportion of category j in the $N(i)$.

kBET

kBET (Buttner et al., 2019) is a statistical test focusing on the local batch label distribution in the k-nearest neighbor graph and compares it with the global batch label distribution to derive a null distribution. For each data, the batch label distribution of its k-nearest neighbors is compared to the expected batch label distribution. After

performing this computation for a subset or all data points, the kBET rejection rate can be calculated, which quantifies the overall batch effects (a low rejection rate implies minor batch effects).

Principal Component Regression

PCR, through PCA, dismantles the original data matrix into principal components (PCs) to primarily assess the variance of this data explained by the batch covariates. By the independences between PCs, the variance of data matrix explained by batch covariate is calculated by the sum of the product of the variance of data matrix explained by each PC_i and variance of each PC_i explained by batch covariates:

$$\mathrm{Var}(C|B) = \sum_{i=1}^{N} \mathrm{Var}(C|PC_i)R^2(PC_i|B)$$

where C is the data matrix, B is the batch covariate.

Isolated Labels

"Isolated labels" is a metric for the identification accuracy of cell labels shared in only few batches. This metric is calculated in two versions: F1 score and ASW score.

Spatial-domain level Multi-batch Integration Metrics

Identifying spatial domains is a task distinct from cell-type clustering, as it heavily relies on the spatial location of data points. In contrast, cell-type clustering focuses more on the inherent properties and similarities amongst data points, without consideration for geographical context. For evaluating the spatial domains identified, we use two additional metrics: Moran's I and Geary's C.

Moran's I

Moran's I measures the level of spatial autocorrelation and serves as an indicator of spatial pattern divergence from randomness. This aids in understanding the spatial continuity within identified regions. Given a set of N cells/spots with their neighboring graph $\boldsymbol{W} = \{w_{ij}\}_{i,j=1}^{N}$ and the corresponding spatial domains $\mathbf{x} = \{x_i\}_{i=1}^{N}$ with its mean $\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$, the Moran's I is calculated as

$$I = \frac{N\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The neighboring graph $\boldsymbol{W}$ is constructed as kNN graph with k values ranging from 1, 10, 20, 30.

Geary's C

Geary's C assesses spatial heterogeneity by measuring dissimilarity between neighboring locations, thus providing insight on the smoothness of the spatial distribution.

By employing these metrics, we can comprehensively evaluate the performance

of spatial domain identification in terms of continuity and smoothness. Given a set of N cells/spots with their neighboring graph $\boldsymbol{W} = \{w_{ij}\}_{i,j=1}^{N}$ and the corresponding spatial domains $\mathbf{x} = \{x_i\}_{i=1}^{N}$, the Geary's C is calculated as

$$C = \frac{(N-1)\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left(x_i - x_j\right)^2}{2\sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\sum_{i=1}^{N}(x_i - \bar{x})^2}$$

The neighboring graph $\boldsymbol{W}$ is constructed as kNN graph with values of k ranging from 1, 10, 20, 30.

Cell-type Deconvolution Metrics

Pearson Correlation Coefficient (PCC)

PCC is a measure of the linear correlation between two variables. It's a value between -1 and 1 where -1 is total negative linear correlation, 0 is no linear correlation, and 1 is total positive linear correlation.

Given two vectors $\mathbf{x}$ and $\mathbf{y}$ of n samples, the Pearson Correlation Coefficient ρ can be calculated with sample means $\bar{x}$ and $\bar{y}$ using the following formula:

$$\rho(\mathbf{x}, \mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

Spearman Correlation Coefficient (SCC)

SCC measures the strength and direction of monotonic relationship between two variables. This is a non-parametric measure and unlike the Pearson correlation, it doesn't require the assumption of normality or linearity. The coefficient is in the range of -1 to 1, where -1 indicates a perfect decreasing monotonic relationship, 0 indicates no monotonic relationship, and 1 indicates a perfect increasing monotonic relationship. Let $\mathbf{x}$ and $\mathbf{y}$ be the rank orders of two vectors of n samples. The Spearman Correlation Coefficient ρ_S can be calculated as:

$$\rho_S(\mathbf{x}, \mathbf{y}) = 1 - \frac{6\sum d_i^2}{n(n^2 - 1)}$$

where d_i represents the difference between the ranks of each observation on x_i and y_i.

Jensen–Shannon Divergence (JSD)

JSD is a method to measure the similarity between two probability distributions. Given two probability distributions P and Q, the Jensen-Shannon Divergence can be calculated as follows:

$$\mathrm{JSD}(P, Q) = \frac{1}{2}\mathrm{KL}(P \parallel M) + \frac{1}{2}\mathrm{KL}(Q \parallel M)$$

where KL is the Kullback-Leibler divergences and $M = \frac{1}{2}(P + Q)$.

Mean Squared Error (MSE)

MSE is a common metric used to evaluate how well a model's predictions match the actual outcomes, where lower values indicate better performance. The MSE is calculated based on true value vector $\mathbf{y} = \{y_i\}_{i=1}^{n}$ and prediction vector $\hat{\mathbf{y}} =$

$\{\hat{y}_i\}_{i=1}^{n}$ using the following formula:

$$\mathrm{MSE}(\mathbf{y}, \hat{\mathbf{y}}) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Synthetic data for validation of deconvolution

The pseudo-SRT data is constructed in the following process: 1. Choose one SRT data and its tissue section and a multi-batches scRNA-seq dataset as reference; 2. Choose one batch of scRNA-seq data for the simulation of SRT gene-expression and the simulation of technical effect between scRNA-seq and SRT data; 3. Generate spatial domains $\mathcal{D}$ for the tissue of SRT data, where we directly use STEP to identify the spatial domains based on the original SRT data, and the mean cell-type composition $E[\boldsymbol{P}_d]$ of each domain d; 4. For each spot, sample number of cells, cell-type composition and corresponding gene expression from uniform distribution, Normal distribution and scRNA-seq dataset, respectively. 5. Perform a spatially local smooth on cell-type composition $\{\boldsymbol{P}_d\}_{d\in\mathcal{D}}$ to simulate the spatially correlation with 2 graph convolutions; 6. Mix the sampled gene expressions and cell-type composition as gene profiles for the spots;

Analyzed Datasets and Analysis Details

Human Colorectal Cancer and Mouse Small Intestine by 10x Visium HD

We downloaded Human Colorectal Cancer and Mouse Small Intestine data from 10x public dataset (https://www. 10xgenomics.com/resources/datasets). Human Colorectal Cancer data with bin size 16μm was used to test. Mouse Small Intestine data with bin size 8μm was tested. STEP was used to analyze these two data with default parameters setting.

Atlas-level Integration scRNA-seq datasets

We downloaded the 6 benchmarking scRNA-seq datasets from GitHub repository of scib-reproducibility (https://github.com/theislab/scib-reproducibility) including 2 simulation datasets (Simulation1, Simulation2) and 4 real datasets (Human Pancreas, Human Lung, Human Immune and Human & Mouse Immune). The Simulation1 dataset contains 12097 cells and 6 batches; Simulation2 dataset contains 19318 cells and 16 batches. Human Pancreas dataset contains 16382 cells and 9 batches; Human Lung dataset contains 32472 cells and 16 donors; Human Immune dataset contains 33056 cells and 10 donors; and Human & Mouse Immune dataset contains 97952 and 23 samples. All the results of compared methods were also obtained from scib-reproducibility. STEP was tested with 2000 HVGs referring to the scib-pipeline and the same parameters-setting across all datasets: 32 modules; hidden dimensions $D = 64, d = 30$; $\beta = 0.01$, and was trained in 400 maximum epochs and 1024 batch-size. All the results of STEP are presented in the Supplementary Figures and Tables.

Human DLFPC dataset by 10x Visium

We obtained 12 space ranger outputs of DLFPC dataset from GitHub repository spatialLIBD (https://github.com/LieberInstitute/HumanPilot) as well as the manual annotations.

For the spatial domain identification of single section, STEP was tested with the parameters listed in (Supplementary Tables 2). The compared methods were all tested referring to the original studies, correspondingly. The complete results are presented on the Supplementary Figures.

For the integration of multiple sections, we set the parameters of STEP as follows: top-3000 HVGs,16 modules, hidden dimensions $D = 64, d = 30$, $\beta = 0.001$, 4 graph convolutional layers. And we trained the STEP with 1000 iterations and graph-sample-size equal to 1. The compared method BASS-Mult was tested with the provided official workflow. The other compared method PRECAST was tested with the default settings. The complete results are presented on Supplementary Figures and Tables.

Mouse Sagittal Anterior and Posterior by 10x Visium

These two sections were obtained from 10x Visium public dataset (https://www.10xgenomics.com/resources/datasets). We set the parameters of STEP as follow: top-3000 HVGs,16 modules, hidden dimensions $D = 40, d = 30$, $\beta = 0.01$, 3 graph convolutional layers. STEP was trained in 800 iterations with graph-sample-size equal to 2.

Mouse Hypothalamus data by MERFISH

We downloaded this data from Dryad (https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248). We conduct the tests for the cell-type level integration and domain level integration on the 5 sections (Bregma distance -0.04 to -0.24) from first animal. Manual annotations for functional regions were obtained from (Li & Zhou, 2022), and all genes were used to perform tests. For the cell-type level integration test, STEP was set as 20 modules, hidden dimensions $D = 40, d = 30$, $\beta = 0.01$ and was trained in 400 epochs. The compared methods were all tested with default parameters. For the test of domain-level integration, STEP was set with additional 4 graph convolutional layers and was trained in 1600 iterations with graph-sample-size equal to 2. In this test, BASS-Mult was tested following the official tutorial and PRECAST was tested with 30 nearest neighbors and all genes.

Furthermore, additional integration of 12 sections (Bregma distance from 0.26 to -0.29) from the 7-th animal in original study were tested by STEP. The complete results are presented in Supplementary Figures.

Mouse medial prefrontal cortex data by STARmap

We downloaded this dataset from STARmap public dataset (https://www.Starmapresources.org/data). We conducted tests on 3 sections: BZ5, BZ9, BZ14, and obtained manual annotations from (Li & Zhou, 2022). For the simplicity of the data, STEP was set to 8 modules, hidden dimensions $D = 40, d = 30$, $\beta = 0.01$ and was trained in 1600 iterations, 2 graph-sample-size. BASS-Mult was tested by the official workflow and PRECAST was tested with all genes and 30 nearest neighbors.

Mouse Liver Atg7 KO Autophagy data by 10x Visium

The reference scRNA-seq data was obtained from Liver Cell Atlas (https://www.livercellatlas.org). For the co-embedding stage, we selected top-2000 HVGs and set STEP with 32 modules, hidden dimensions $D = 64, d = 30$; $\beta = 0.01$. In this stage, STEP was trained in 400 maximum epochs and 1024 batch-size. The spatial domains were identified alone without using co-embedding and ST-decoder. For the deconvolution stage, STEP was set to run 2000 iterations based on the identified spatial domains. All the results are presented in the Supplementary Figures.

Data availability

All relevant data supporting the key findings of this study are available within the article and its Supplementary Information files. The datasets used for the analysis in this study are available from the following sources: (1) Human Colorectal Cancer and Mouse Small Intestine data from 10x Visium HD at https://www.10xgenomics.com/resources/datasets;(2) six benchmarking scRNA-seq datasets from scib-reproducibility GitHub repository at https://github.com/theislab/scib-reproducibility, including Simulation1, Simulation2, Human Pancreas, Human Lung, Human Immune, and Human & Mouse Immune datasets; (3) 12 space ranger outputs of DLFPC dataset from spatialLIBD GitHub repository at https://github.com/LieberInstitute/HumanPilot; (4) Mouse Sagittal Anterior and Posterior sections data from 10x Visium public dataset at https://www.10xgenomics.com/resources/datasets; (5) Mouse Hypothalamus data by MERFISH from Dryad at https://datadryad.org/stash/dataset/doi:10.5061/dryad.8t8s248; (6) Mouse medial prefrontal cortex data by STARmap at https://www.starmapresources.org/data; (7) Reference scRNA-seq data for Mouse Liver from Liver Cell Atlas at https://www.livercellatlas.org.

The in-house data, Mouse Liver Atg7 KO Autophagy data by Visium, has been uploaded on GEO.

Code availability

The open-source package of STEP is available at GitHub repository: https://github.com/SGGb0nd/step. The tutorial and documentation is hosted in https://sggb0nd.github.io/step.

Author contributions

X.X. and Z.L. designed the study. L.L. designed the model and algorithm. L.L. developed STEP software and related documentations. L.L. and X.X. and Z.L. wrote the manuscript. L.L. and X.X. performed data analysis. Y.L. and X.Y. provide the mouse liver ATG7 KO data, and performed annotation as well as the interpretation for the analysis results. All authors read and approved the manuscript.