Deciphering spatial domains from spatial multi-omics with SpatialGlue

Yahui Long[1], Kok Siong Ang[1], Raman Sethi[1], Sha Liao[2,3], Yang Heng[2,3], Lynn van Olst[4], Shuchen Ye[1], Chengwei Zhong[1], Hang Xu[1], Di Zhang[5], Immanuel Kwok[1], Nazihah Husna[1], Min Jian[2,6], Lai Guan Ng[1], Ao Chen[2,3,7], Nicholas RJ Gascoigne[8,9,10], David Gate[4], Rong Fan[5], Xun Xu[2], Jinmiao Chen[1,9,10*]

[1]Singapore Immunology Network (SIgN), Agency for Science, Technology and Research (A*STAR), 8A Biomedical Grove, Immunos Building, Level 3, Singapore 138648, Singapore

[2]BGI-Shenzhen, Shenzhen, Guangdong, China

[3]BGI Research-Southwest, BGI, Chongqing 401329, China

[4]The Ken & Ruth Davee Department of Neurology, Northwestern University Feinberg School of Medicine, Chicago, IL, USA.

[5]Department of Biomedical Engineering, Yale University, New Haven, CT, USA.

[6]BGI Research Asia-Pacific, BGI, Singapore 138567, Singapore

[7]JFL-BGI STOmics Center, Jinfeng Laboratory, Chongqing 401329, China

[8]Immunology Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, Singapore 117545, Singapore

[9]Department of Microbiology and Immunology, Yong Loo Lin School of Medicine, National University of Singapore, 5 Science Drive 2, Singapore 117545, Singapore

[10]Cancer Translational Research Programme, Yong Loo Lin School of Medicine, National University of Singapore, Singapore, Singapore

*Corresponding author. Email: chen_jinmiao@immunol.a-star.edu.sg

Abstract

Integration of multiple data modalities in a spatially informed manner remains an unmet need for exploiting spatial multi-omics data. Here, we introduce SpatialGlue, a novel graph neural network with dual-attention mechanism, to decipher spatial domains by intra-omics integration of spatial location and omics measurement followed by cross-omics integration. We demonstrate that SpatialGlue can more accurately resolve spatial domains at a higher resolution across different tissue types and technology platforms, to enable biological insights into cross-modality spatial correlations.

Key words: Spatial multi-omics, Cross-omics integration, Deep learning, Graph neural networks, Dual attention

Main

Spatial transcriptomics is the next major development in analyzing biological samples since the advent of single-cell transcriptomics. Currently, spatial technologies are expanding to spatial multi-omics with the simultaneous profiling of different omics on a single tissue section. These technologies can be roughly divided into two categories, sequencing-based and imaging-based. Sequencing-based techniques include DBiT-seq [1], spatial-CITE-seq [2], spatial-ATAC-RNA-seq and CUT&Tag-RNA-seq [3], SPOTS [4], SM-Omics [5], Stereo-CITE-seq [6], spatial RNA-TCR-seq [7], and 10x Genomics Xenium [8],while imaging-based techniques include DNA seqFISH+ [9], DNA-MERFISH based DNA and RNA profiling [10], MERSCOPE [11], and Nanostring CosMx [12]. To fully utilize spatial multi-omics data to construct a coherent picture of the tissue under study, spatially aware integration of heterogeneous data modalities is required. Such multi-omics data integration poses a significant challenge as different modalities have feature counts that can vary enormously (e.g., number of proteins vs transcripts measured) and possess different statistical distributions. This challenge is deepened when integrating spatial information with feature counts within each data modality. To our knowledge, there is no tool designed specifically for spatial multi-omics acquired from the same tissue section. Existing methods are either unimodal or do not use spatial information, except for one tool with functionality for spatial multi-omics integration, MEFISTO, which has only been previously demonstrated on single-cell multi-omics or spatial transcriptomics separately. Other tools such as STAGATE [13], SpaGCN [14] and GraphST [15] target spatial single omics analysis, while Seurat WNN [16], MOFA+ [17], totalVI [18], MultiVI [19], scMM [20], and StabMap [21] perform multi-omics data integration without employing spatial information. There is consequently a great need for spatially aware cross-omics integration methods specifically designed for spatial multi-omics.

Here we introduce SpatialGlue, a graph neural network (GNN) based deep learning model that performs spatial multi-omics data analysis (Figure 1a). The input data to SpatialGlue can be feature matrices of segmented cells or capture locations (beads, voxels, pixels, bins, or spots), with accompanying spatial coordinates. We refer to the cells and the capture locations as spots here after for brevity and not to restrict SpatialGlue to any specific technological platform or resolution. Conceptually, SpatialGlue employs a dual attention mechanism to achieve data integration on two levels, within-modality spatial information and measurement feature integration first, and then between-modality multi-omics integration. SpatialGlue first learns a low dimension embedding within each modality using spatial and omics data. Within each modality, SpatialGlue constructs a spatial proximity graph and a feature similarity graph which are used separately to encode the pre-processed expression data into a common low dimension embedding space. Here the spatial proximity graph captures spatial relationships between spots, while the feature graph captures feature similarities. These constructed graphs possess unique semantic information that can be integrated to better capture cellular heterogeneity. However, the different graphs can contribute differential importance to each spot, posing a challenge to capture this difference. Therefore, we adopted a within-modality attention aggregation layer to adaptively integrate the spatial and feature graph-specific representations and derive modality-specific representations. Specifically, the model learns graph-specific weights to assign importance to each graph. Similarly, the different omics modalities can have distinct and complementary contributions to each spot. Thus, we further designed a between-modality attention aggregation layer that learns

modality-specific importance weights and adaptively integrates the modality-specific representations to generate the final cross-modality integrated latent representation. The learned weights illustrate the contribution of each modality to the learned latent representation of each spot and consequently the demarcation of different spatial domains or cell types. We believe this approach enables more accurate integration than summation or concatenation of the feature matrices. We validated the importance of attention and other components with a series of ablation studies (See Supplementary file). After obtaining SpatialGlue's integrated multi-omics representation, we can then employ clustering to identify biologically relevant spatial domains which consist of cells that are coherent spatially and across the measured omics. Such spatial domains can range from local clusters of distinct cell states to functionally distinct anatomical structures.

We first benchmarked SpatialGlue with competing methods using simulation data and experimentally acquired data with ground truth labels. With ground truth available, we can assess performance with supervised metrics, namely homogeneity, mutual information, v measure, AMI, NMI, and ARI. We generated a set of simulated data consisting of two modalities that together contained the information of the ground truth (Figure 1b, left). The modalities were designed to simulate the transcriptome and proteome, respectively, with the first modality following the zero-inflated negative binomial (ZINB) distribution and the second following the negative binomial (NB) distribution (Figure 1c). For comparison, we tested seven competing methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, and StabMap, alongside SpatialGlue. Visually, SpatialGlue was able to clearly recover all four spatial factors to closely match the ground truth (Figure 1b). Seurat and MEFISTO were able to clearly recover two factors (factors 2 and 4 for Seurat, 3 and 4 for MEFISTO). Other methods were able to recover some of the factors but with much higher levels of noise (factor 2 for totalVI, 1 and 2 for MOFA+, MultiVI, and scMM, 2 and 3 for StabMap). The metrics confirmed the visuals with SpatialGlue scoring top in all metrics, followed by Seurat and MEFISTO (Figure 1e). We further tested all methods with four more datasets generated with modified distribution parameters (Suppl. Figure S1-3) and measured the performance with the same metrics, summarizing the results with box plots (Figure 1f). Again, SpatialGlue performed the best with little variance between different datasets. Lastly, we have also demonstrated on simulation data that the SpatialGlue framework is extensible to three or more modalities (Suppl. Figure S4).

For the second example, we benchmarked SpatialGlue and the same competing methods with an in-house human lymph node dataset generated using 10x Genomics Visium RNA and protein co-profiling technology (section A1). Here we used the H&E based annotation as the ground truth (Figure 1g). In the annotation, the major structures include the pericapsular adipose tissue and capsule that form the outer layers of the bulb, while the cortex and medulla (cords and vessels) form the core internal structures. For comparison, we also plotted the single modality PCA-based clustering of RNA and protein (Figure 1h, left). All the methods were able to isolate the paracortex (T cell zone, SpatialGlue cluster 1) that more resembled the RNA and protein specific modalities than the H&E annotation, which is unsurprising because T cells can be better identified by protein and gene markers such CD8A, CD3E, and CCR7 (Suppl. Figure S10). The methods were also unable to differentiate capsule layers from the pericapsular adipose tissue, which were also not well captured in the RNA and protein modalities either. Among the tested

methods, SpatialGlue, Seurat, totalVI, and MOFA+ were able to identify the follicle regions while MultiVI, scMM, MEFISTO, and StabMap could not. The hilum, which normally accumulates fat, is only visible in the RNA modality and only MOFA+ and SpatialGlue could separate it from the pericapsular layer. To assess performance quantitatively, we employed both unsupervised and supervised metrics. We first used the unsupervised Moran's *I* score and Jaccard Similarity to assess spatial autocorrelation of clusters and preservation of distance in the joint latent space, respectively. The Moran's *I* score was computed for each cluster and plotted as a box plot for each method. SpatialGlue outperformed all other methods with a median score of 0.62 (Figure 1i). We computed a Jaccard Similarity score to quantify the overlap of neighbor sets between the joint space and each modality. Summed together, the total Jaccard Similarity of SpatialGlue also outperformed the other methods with MOFA+ as a close second (Figure 1j). For the supervised metrics computed with respect to the ground truth, SpatialGlue likewise outperformed all other methods with 6 clusters (Suppl. Figure S8c). We further generated different numbers of clusters and the resulting box plots of supervised metrics showed results stability regardless of clustering resolution (Figure 1k). To ensure that the results were not predicated on a specific tissue section, we applied the same methods to another human lymph node section (D1). With this data, SpatialGlue showed comparable scores with 6 clusters, but achieved more stable performance across different clustering resolutions than other methods (Figure S8).

Next, we applied SpatialGlue to mouse brain epigenome-transcriptome datasets to showcase its ability to resolve spatial domains at a higher resolution than methods used in the original study. We first tested SpatialGlue on a P22 mouse brain coronal section dataset acquired using spatial ATAC-RNA-seq [3] to measure mRNA and open chromatin regions. We employed the Allen brain atlas reference to annotate anatomical regions such as the cortex layers (ctx), genu of corpus callosum (ccg), lateral septal nucleus (ls), and nucleus accumbens (acb) (Figure 2a). For benchmarking, we tested SpatialGlue against Seurat, MultiVI, MOFA+, scMM, and StabMap. We did not include MEFISTO and totalVI because we could not finish running MEFISTO within 12 hours, and totalVI was designed only for CITE-seq. We first visualized the individual modalities (Figure 2b), where we see that they captured various regions with differing accuracy. While both modalities captured the lateral ventricle (vl) and the lateral preoptic area (lpo), the RNA modality clearly captured the ccg and olfactory limb of the anterior commissure (aco) but was unable to differentiate the ctx layers. Meanwhile, the ATAC modality was able to isolate the caudoputamen (cp) as well as some of the ctx layers. SpatialGlue captured all of the aforementioned anatomical regions (2-acb, 4-cp/13-cp, 9-vl, 11-ccg/aco, 12-ls, 18-lpo) and produced better defined layers in the ctx and anterior cingulate area (aca) regions. Notably, SpatialGlue was able to differentiate more ctx layers than all other methods including the original analysis by Zhang et al. Seurat was able to capture the vl, acb, cp, and ctx layers, making it the second-best method. While the other methods could only capture the ccg and a few of the other structures. In general, the outputs of competing methods presented more noise than SpatialGlue, which is quantitatively confirmed by the Moran's *I* score (Figure 2c). For the Jaccard Similarity, SpatialGlue again ranked top (Figure 2d). We next examined the cross-modality and intra-modality weights learned by SpatialGlue in the aggregation layers. These weights denote the contribution of individual modality's features and spatial information towards the integrated output (Figure 2e, Suppl. Figure S12c). For the cross-modality weights, the RNA modality better segregated the ccg/aco region and thus was

assigned a heavier weight. While for the ctx and vl, the ATAC modality showed more contribution and thus a heavier weight was assigned.

We extended the analysis to another P22 mouse brain dataset of a highly similar coronal section but with RNA-seq and CUT&Tag (H3K27ac histone modification) modalities. This dataset also does not have an annotated ground truth; we therefore again used the Allen brain atlas reference for annotation. In this dataset, SpatialGlue captured the major structures of the ctx layers (clusters 1,2,5,6,12), 8-aca, 10-ccg/aco, cp (7,14), vl (9,16), 3-ls, and 4-acb (Figure 2f). All other methods were unable to clearly capture many structures such as the acb and ls. The output of SpatialGlue also had the least noise, which was also reflected in Moran's *I* score (Figure 2g). For the Jaccard Similarity, SpatialGlue achieved the highest score, highlighting that SpatialGlue's integrated output was able to best preserve the between-spot distance from the original individual data modalities (Figure 2h). We also examined the modality weights for the contribution of the different modalities towards each cluster (Figure 2i). For most clusters, the histone modification modality made similar or greater contribution. Most notable is the vl structure (cluster 9, 16), which is strongly visible in the histone modification modality plot. To ensure that the results were not contingent on dataset selection, we again tested on two other P22 mouse brain dataset with RNA-seq and CUT&Tag (H3K4me3 and H3K27me3 histone modification) modalities. SpatialGlue was again the top method in both Moran's *I* score and Jaccard Similarity for these datasets (Figure S14,15).

We further analyzed the DEGs of each cluster (Figure 2j) and found known markers for the different brain regions such as myelin related genes, Tspan2, Cldn11, and Ugt8a, expressed in the post-natal developing corpus callosum (10-ccg/aco), and Olfm1, Cux2, Rorb in the cortex layers. We next examined the differential expressed peaks in the H3K27ac histone modification modality (Figure 2k), where we found strong peaks in the clusters 12-ctx, 10-ccg/aco, 4-acb, and 7-cp. Finally, we plotted the peak-to-gene links heatmap (Figure 2l). Here, there are two major groups appearing in both data modalities, the first primarily consisting of acb/cp structures (4-acb, 7-cp, and 14-cp), and the second of ctx-related clusters (6-ctx, 11-ctx, and 12-ctx). This illustrates SpatialGlue's success in combining information from both modalities into the latent space to enable biologically relevant clusters. We believe such information combination has also contributed to the detection of the 4 cortical layers (cluster 5, 12, 1, 6). Within the cortex layers 5 and 6 (cluster 6), Tle4, Fezf2, Foxp2, and Ntsr1 have been reported in literature as markers. However, we only found Tle4 and Fezf2 expression to spatially coincide with the cluster. Conversely, Ntsr1's gene activity score inferred from the histone marks matched the cluster spatially (Suppl. Figure S13). This illustrated the different information within each modality that SpatialGlue can leverage to better demarcate different spatial domains.

Lastly, we demonstrated that SpatialGlue is broadly applicable to a wide spectrum of technology platforms by further applying it to Stereo-CITE-seq and SPOTS acquired data. The Stereo-CITE-seq [6] was used to analyze a mouse thymus section, capturing mRNA and protein at sub-cellular resolution (Figure 3a left). The thymus is a small gland surrounded by a capsule of fibers and collagen). It is divided into two lobes connected by a connective isthmus with each lobe being broadly divided into a central medulla surrounded by an outer cortex layer. In each data modality, broad outlines of the medulla regions and the surrounding cortex could be seen (Figure

3a). We tested eight methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. MultiVI and StabMap were unable to find coherent clusters that resembled the medulla and cortex regions within the thymus. This is clearly reflected in the Moran's *I* score and Jaccard Similarity with these two methods scoring the lowest (Figure 3b,c). Seurat, totalVI, scMM, and SpatialGlue were more successful in capturing the internal structures by separating the medulla from the cortex, with SpatialGlue and scMM better demarcating the CMJ and the inner, middle, and outer cortex (clusters 2, 3, 4, 5). Overall, SpatialGlue scored the highest in Jaccard Similarity and second in Moran's *I* score. This superior performance was further replicated with three additional mouse thymus sections (Suppl. Figures S17, 18, 19). For most clusters, the RNA modality made greater contributions than the protein (Figure 3d). But for the inner cortex (cluster 3), the protein modality contributed more and its contribution to distinguishing the inner cortex is visible in the protein modality spatial plot (Figure 3a left).

Finally, we benchmarked SpatialGlue's capabilities with murine spleen spatial profiling data consisting of protein and transcript measurements [4]. The spleen is an important organ within the lymphatic system with functions including B cell maturation in germinal centers formed within B cell follicles (Figure 3e). These are complex structures with an array of immune cells present. The data was generated using SPOTS that employs the 10x Genomics Visium technology to capture whole transcriptomes and extracellular proteins via polyadenylated antibody-derived tag-conjugated (ADT-conjugated) antibodies. The protein detection panel used for this experiment was designed to detect the surface markers of B cells, T cells, and macrophages which are well represented in the spleen. After preprocessing, we performed clustering of each data modality and plotted the clusters on the tissue slide to examine their correspondence between modalities (Figure 3f, left). The clusters clearly did not align, indicating that each modality possessed different information content (Suppl. Figure S24f, g). Using the protein markers and DEGs, clusters of spots enriched with B cells, T cells, and macrophage subsets were annotated [22–24]. Specifically, we identified macrophage subsets (RpMΦ, MZMΦ, MMMΦ) that were not annotated in the original study. We tested Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue (Figure 3f). MultiVI and StabMap did not capture coherent clusters. This was also reflected in their Moran's *I* scores and Jaccard Similarity (Figure 3g, h). The remaining methods captured clusters with similar Moran's *I* score but SpatialGlue scored the highest in Jaccard Similarity. We then examined SpatialGlue's learned modality weights (Figure 3i). The protein modality made the bigger contribution to the MMMΦ cluster, which was mainly found in the protein modality plot. Conversely, SpatialGlue relied more on the RNA modality to capture the T cell cluster. To verify SpatialGlue performance, we used another SPOTS acquired dataset of a murine spleen section as replicate. Here, SpatialGlue achieved comparable or greater Moran's *I* score than baseline methods and scored the highest in terms of Jaccard Similarity (Figure S25e, f).

To annotate the clusters found with SpatialGlue, we visualized the cell types' protein markers (Figure 3j, k) and RNA expression of select markers (Figure 3l). Within the white pulp zone, the T cell spots were known to concentrated in small clusters known as T cell zones while the B cell enriched spots were mainly found in areas adjacent to the T cell clusters. The RpMΦ markers were unsurprisingly the strongest in the red pulp zone, being easily identifiable with markers like F4_80 and CD163. To differentiate MZMΦ and MMMΦ, the RNA expression of Cd209a (MZMΦ) and Siglec1 (MMMΦ) were used to guide the annotation.

From the cluster and marker visualization, we observed cell types which were spatially adjoining. Thus, we quantitated the spatial relationship by computing neighborhood enrichment (Suppl. Figure S24d) and co-occurrence scores with respect to distance from the T and B cell perspectives (Suppl. Figure S24e). In general, we observed neighborhood enrichments that matched known biology such as the high correlation between the B and T cells, indicating that they are most likely to be found together at the closest distance. This was followed by MMMΦ which surrounded T and B cell clusters in the white zone. These reflected the layers of cell types that form the follicles and their surroundings. Between the macrophages, we see positive correlation between RpMΦ and MZMΦ, and MZMΦ and MMMΦ. This is a result of the red pulp (RpMΦ) forming the spleen's outer layer and the MZMΦ being positioned within the marginal zone surrounding white pulp which in turn was enriched with MMMΦ.

SpatialGlue is a novel deep learning model incorporating graph neural networks with dual attention mechanisms that enables integration of multi-omics data in a spatially aware manner. With the presented examples, we demonstrated SpatialGlue's ability to effectively integrate multiple data modalities with their respective spatial context to reveal histologically relevant structures of tissue samples. Furthermore, our quantitative benchmarking demonstrated that SpatialGlue exhibits superior performance to 10 state-of-the-art unimodal and non-spatial methods on 5 simulated data and 12 real datasets, highlighting the importance of spatial information and cross-omics integration. We also demonstrated SpatialGlue's ability to resolve finer grained tissue structures, which can facilitate novel biological findings in future studies. For example, its application to mouse brain epigenome-transcriptome data revealed finer cortical layers compared to the original study, which can allow further investigation of gene regulation at a higher spatial resolution. Our examples also spanned four different tissue types and four technology platforms to show its broad applicability. Despite having demonstrated its application only on sequencing-based spatial omics data, its design allows seamless extension to image-based omics data from technology platforms like 10x Genomics Xenium and Nanostring CosMx, exhibiting a technology-agnostic nature.

As a graph neural network (GNN)-based method, SpatialGlue bears such similarity to other GNN methods such as GraphST and SpaGCN. Naturally, as a method tailored for spatial multi-omics, it is different as it is explicitly designed to take in multiple data modalities as input and employ attention to integrate data, as opposed to concatenation at data preprocessing. Unlike other existing multi-modal methods such as Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, and StabMap, our model is spatially informed and it adopts attention mechanism to adaptively learn the relative importance between omics modalities, and between spatial location and omics feature within each modality.

We also designed SpatialGlue to be computation resource efficient and thus relevant as data sizes increase. The largest dataset tested contained 9,752 spots (spatial-epigenome-transcriptome mouse brain), and it required about 5 mins of wall-clock time on a server equipped with an Intel Core i7-8665U CPU and NVIDIA RTX A6000 GPU. It scales well with number of features and cells/spots (as shown in Suppl. Figure S7g). Therefore, we believe SpatialGlue will be an invaluable analysis tool for present and future spatial multi-omics data. Most technologies can produce accompanying imaging data such as H&E, which contains essential information of

the cell and tissue morphology. The integration of image modality is currently lacking in SpatialGlue. In future, we plan to extend SpatialGlue to incorporate image data at either the intra- or inter-modality attention aggregation layer. We also plan to extend SpatialGlue's functionality with integration of multi-omics data acquired from adjacent tissue sections.

Methods

Data

Human lymph node dataset For spatial transcriptomics analysis of human tissues, two sequential sections of 5 µm thickness were utilized from formalin-fixed, paraffin-embedded (FFPE) lymph node. The sections underwent spatial transcriptomic library construction using CytAssist Visium platform (10x Genomics). Initially, sections were stained with Hematoxylin and Eosin (H&E) following the protocol outlined in the Visium CytAssist guide for FFPE samples, which includes steps for deparaffinization, staining, imaging, and decrosslinking (Reference: CG000658, 10x Genomics, CA, USA). Imaging was performed with a 20x objective on an EVOS M7000 microscope (Thermo).

Following imaging, spatial gene expression libraries were prepared utilizing probe-based methods, along with spatial protein expression libraries as per the guidelines provided in the Visium CytAssist Reagent Kits manual (Reference: CG000494, 10x Genomics, CA, USA). We employed the Visium Human Transcriptome Probe Set version 2.0 for RNA transcript detection, along with the Human FFPE Immune Profiling Panel, which includes a 35-plex CytAssist Panel of antibodies, both intracellular and extracellular, sourced from BioLegend and Abcam for protein detection. This panel also comprises four isotype controls. Antibody signals were normalized to isotype controls.

Libraries were sequenced on an Illumina NovaSeq S2 PE50 platform, allocating 2000 million reads per lane at the NUSeq Core Facility, Northwestern University. The resultant FASTQ files were processed using the spaceranger-2.1.0 software, referencing the GRCh38 human genome (GENCODE v32/Ensembl 98). For precise anatomical context, we conducted manual annotation of the lymph node structures, utilizing the high-resolution images captured by the EVOS M7000 microscope within the Loupe Browser software (10x Genomics, CA, USA).

Spatial-epigenome-transcriptome mouse brain dataset Brain tissue sections from a juvenile (P22) mouse was analyzed for the epigenome and transcriptome using spatial-ATAC-RNA-seq and CUT&Tag-RNA-seq by Zhang et al. [3]. Microfluidic barcoding was used to capture spatial location and combined with in situ Tn5 transposition chemistry to capture chromatin accessibility. We used four datasets, one spatial-ATAC-RNA-seq dataset and three spatial CUT&Tag-RNA-seq datasets. The number of pixels ranged from 9,215 to 9,752, the number of genes ranged from 22,731 to 25,881, and the number of peaks ranges from 35,270 to 121,068.

To preprocess the transcriptomic data, pixels expressing fewer than 200 genes and genes expressed fewer than 200 pixels were filtered out. Next, the gene expression counts were log-transformed and normalized by library size via the SCANPY package [25]. The top 3,000 highly variable genes (HVGs) were selected and used as input to PCA for dimensionality reduction. For

consistency with the chromatin peak data, the first 50 principal components were retained and used as input to the encoder. For the chromatin peak data, we used LSI (latent semantic indexing) to reduce the raw chromatin peak counts data to 50 dimensions.

Stereo-CITE-seq mouse thymus dataset Murine thymus tissue samples were investigated with Stereo-CITE-seq for spatial multi-omics by Liao et al.[6]. For our study, we employed data from four sections. The number of bins ranges from 4,228 to 4,697, the number of genes ranges from 23,221 to 23,960, and the sample includes 19 or 51 proteins. For the transcriptomic data, we first filtered out genes expressed in fewer than 10 bins and bins with fewer than 80 gene expressed. The filtered gene expression counts were next log-transformed and normalized by library size via the SCANPY package [25]. Finally, to reduce the dimensionality of the data, the top 3,000 highly variable genes (HVGs) were selected and used as input for PCA. To ensure a consistent input dimension with the ADT data, the first 22 principal components were retained and used as the input of the encoder. For the ADT data, we first filter out proteins expressed in fewer than 50 bins, resulting 22 proteins retained. The protein expression counts were then normalized using CLR (Centered Log Ratio) across each bin. PCA was then performed on the normalized data, and all 22 principal components were used as the input of the encoder.

SPOTS mouse spleen dataset Ben-Chetrit et al. [4] processed fresh frozen mouse spleen tissue samples and analyzed them using the 10x Genomics Visium system supplemented with DNA-barcoded antibody staining. The antibodies (poly(adenylated) antibody-derived tags (ADTs)) enabled protein measurement alongside the transcriptome profiling by 10x Genomics Visium. The panel of 21 ADTs was designed to capture the markers of immune cells found in the spleen, including B cells, T cells, and macrophages. We employed two datasets (replicate 1 and 2) from the original study. The data contained 2,568 and 2,768 spots for replicates 1 and 2, respectively, with 32,285 genes captured per spot. For data pre-processing, we first filtered out genes expressed in fewer than 10 spots. The filtered gene expression counts were then log-transformed and normalized by library size using the SCANPY package [25]. Finally, the top 3,000 HVGs were selected and used as input for PCA. We used the first 21 principal components as the input of the encoder to ensure a consistent input dimension with the ADT data. For the ADT data, we applied CLR normalization to the raw protein expression counts. PCA was then performed on the normalized data and the top 21 principal components were used as input to the encoder.

The SpatialGlue framework

SpatialGlue is a novel graph-based model with dual-attention mechanism that aims to learn a unified representation by fully exploiting the spatial location information and expression data from different omics modalities. Within each modality, SpatialGlue first learns a modality-specific representation using both spatial and omics data. Subsequently, it synthesizes an integrated cross-modality representation by aggregating these modality-specific representations. Compared to cross-omics integration first followed by spatial integration, our approach allows us to capture modality-specific spatial correlations between spots and integrate the spatial information in a modality-specific manner.

We first consider a spatial multi-omics dataset with two different omics modalities, each with a distinct feature set $X_1 \in R^{N \times d_1}$ and $X_2 \in R^{N \times d_2}$. N denotes the number of spots in the tissue. d_1

and d_2 are the numbers of features for two omics modalities, respectively. For example, in spatial-epigenome-transcriptome, X_1 and X_2 refer to the sets of genes and chromatin regions respectively, while in Stereo-CITE-seq, X_1 and X_2 refer to the sets of genes and proteins, respectively. The primary objective of spatial multi-omics data integration is to learn a mapping function that can project the original individual modality data into a uniform latent space and then integrate the resulting representations. As shown in Figure 1a, the SpatialGlue framework consists of four major modules: (1) Modality-specific GCN encoder, (2) Within-Modality attention aggregation layer, (3) Between-Modality attention aggregation layer, and (4) Modality-specific GCN decoder. The details of each module are described next. Notably, here we demonstrate the SpatialGlue framework with two modalities. Benefiting from the modular design, SpatialGlue readily extends to spatial multi-omics data with more than two modalities.

Construction of neighbor graph

Assuming spots that are spatially adjacent in a tissue usually have similar cell types or cell states, we convert the spatial information to an undirected neighbor graph $G_s = (V, E)$ with V denoting the set of N spots and E denoting the set of connected edges between spots. Let $A_s \in R^{N \times N}$ be the adjacent matrix of graph G_s, where $A_s(i, j) = 1$ if and only if the Euclidean distance between spots i and j is less than specific neighbor number r, otherwise 0. In our examples, we select the top $r = 3$ nearest spots as neighbors of a given spot for all datasets according to experimental results.

In a complex tissue sample, it is possible for spots with the same cell types/states to be spatially non-adjacent to each other, or even far away. To capture the proximity of such spots in a latent space, we explicitly model the relationship between them using a feature graph. Specifically, we apply the *k*-nearest neighbor algorithm (KNN) on the PCA embeddings and construct the feature graph $G_f^m = (V^m, E^m)$, where V^m and E^m denote the sets of N spots and connected edges between spots in the $m \in \{1,2\}$-th modality, respectively. For a given spot, we choose the top k nearest spots as its neighbors. By default, we set k to 20 for all datasets. We use $A_f^m \in R^{N \times N}$ to denote the adjacency matrix of the feature graph G_f^m. If spot $j \in V^m$ is the neighbor of spot $i \in V^m$, then $A_f^m(i, j) = 1$, otherwise 0.

Graph convolutional encoder for individual modality

Each modality (e.g., mRNA or protein) contains a unique feature distribution. To encode each modality in a low dimension embedding space, we use the graph convolution network (GCN) [26], an unsupervised deep graph network, as the encoder of our framework. The main advantage of GCNs is that it can capture the cell expression patterns and neighborhood microenvironment while preserving the high-level global patterns. For each modality, using the pre-processed features as inputs, we separately implement a GCN-encoder on the spatial adjacency graph G_s and the feature graph G_f to learn graph-specific representations H. These two neighbor graphs reflect distinct topological semantic relationships between spots. The semantic information in the spatial graph denotes the physical proximity between spots while that in the feature graph denotes the phenotypic proximity of spots which have the same cell types/states but are spatially non-adjacent to each other. This enables the encoder to capture different local patterns and

dependencies of each spot by iteratively aggregating the representations from its neighbors. Specifically, the l-th ($l \in \{1,2,\dots,L-1,L\}$) layer representation in the encoder are formulated as follows:

$$H_{s1}^{l} = \sigma(\tilde{A}_s H_{s1}^{l-1} W_{e1}^{l-1} + b_{e1}^{l-1}), \quad (1)$$

$$H_{f1}^{l} = \sigma(\tilde{A}_f^1 H_{f1}^{l-1} W_{e1}^{l-1} + b_{e1}^{l-1}), \quad (2)$$

$$H_{s2}^{l} = \sigma(\tilde{A}_s H_{s2}^{l-1} W_{e2}^{l-1} + b_{e2}^{l-1}), \quad (3)$$

$$H_{f2}^{l} = \sigma(\tilde{A}_f^2 H_{f2}^{l-1} W_{e2}^{l-1} + b_{e2}^{l-1}), \quad (4)$$

where $\tilde{A} = D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$ represents the normalized adjacency matrix of specific graph and D is a diagonal matrix with diagonal elements being $D_{ii} = \sum_{j=1}^{N} A_{ij}$. In particular, $\tilde{A}_s$, $\tilde{A}_f^1$, and $\tilde{A}_f^2$ are the corresponding normalized adjacency matrices of the spatial graph, the feature graphs of modalities 1 and 2, respectively. $W_{e\cdot}$, and $b_{e\cdot}$ denote a trainable weight matrix and a bias vector, respectively. $\sigma(\cdot)$ is a nonlinear activation function such as the ReLU (Rectified Linear Unit). $H_{\cdot}^{l}$ denotes the l-th layer output representation, and $H_{s1}^{0} = H_{f1}^{0}$ and $H_{s2}^{0} = H_{f2}^{0}$ are set as the input PCA embeddings of the original features X_1 and X_2 respectively. We also specify $H_{\cdot}^{L} \in R^{d_3}$, the output at the L-th layer, as the final latent representation of the encoder with d_3 as the hidden dimension. H_{sm} and H_{fm} represent the latent representations derived from the spatial and feature graphs within modality m, respectively.

Within-Modality attention aggregation layer

For an individual modality, taking its pre-processed features and two graphs (i.e., spatial and feature graphs) as input, we can derive two graph-specific spot representations via the graph convolutional encoder, such as H_{s1} and H_{f1}. To integrate the graph-specific representations, we design a Within-Modality attention aggregation layer following the encoder such that its output representation preserves expression similarity and spatial proximity. Given that different neighbor graphs can provide unique semantic information for each spot (as mentioned above), the aggregation layer is designed to integrate graph-specific representations in an adaptive manner by capturing the importance of each graph. As a result, we derive a modality-specific representation for each modality. Specifically, for a given spot i, we first subject its graph-specific representation h_i^t to a linear transformation (i.e., a fully connected neural network), and then evaluate the importance of each graph by the similarity of the transformed representation and a trainable weight vector q. Formally, the attention coefficient e_i^t, representing the importance of graph t to the spot i, is calculated by:

$$e_i^t = q^T \cdot \tanh\left(W_i^{intra} h_i^t + b_i^{intra}\right), \quad (5)$$

where W_i^{intra} and b_i^{intra} are the trainable weight matrix and bias vector, respectively. To reduce the number of parameters in the model, all the trainable parameters are shared by the different graph-specific representations within each modality. To make the attention coefficient comparable

across different graphs, a softmax function is applied to the attention coefficient to derive attention score α_i^t.

$$\alpha_i^t = \frac{\exp(e_i^t)}{\sum_{t=1}^{T} \exp(e_i^t)}, \quad (6)$$

where T denotes the number of neighbor graphs (set to 2 here). α_i^t represents the semantic contribution of the t-th neighbor graph to the representation of spot i. A higher value of α_i^t means greater contribution.

Subsequently, the final representation Y^m in the m-th modality can be generated by aggregating graph-specific representations according to their attention scores:

$$y_i^m = \sum_{t=1}^{T} \alpha_i^t \cdot h_i^t \quad (7)$$

such that $y_i^m \in R^{d_3}$ preserves the raw spot expressions, spot expression similarity, and spatial proximity within modality m.

Between-Modality attention aggregation layer

Each individual omics modality provides a partial view of a complex tissue sample, thus requiring an integrated analysis to obtain a comprehensive picture. These views can contain both complementary and contradictory elements, and thus different importance should be assigned to each modality to achieve coherent data integration. Here we use a Between-Modality attention aggregation layer to adaptively integrate the different data modalities. This attention aggregation layer will focus on the more important omics modality by assigning greater weight values to the corresponding representation. Like the Within-Modality layer, we first learn the importance of modality m by calculating the following coefficient g_i^m:

$$g_i^m = v^T \cdot \tanh(W_i^{inter} y_i^m + b_i^{inter}), \quad (8)$$

where g_i^m is attention coefficient that represents the importance of the modality m to the representation of spot i. W^{inter}, b^{inter}, and v are learnable weight and bias variables, respectively. Similarly, we further normalize the attention coefficients using the softmax function:

$$\beta_i^m = \frac{\exp(g_i^m)}{\sum_{m=1}^{M} \exp(g_i^m)}, \quad (9)$$

where β_i^m is the normalized attention score that represents the contribution of the modality m to the representation of spot i. M is the number of modalities.

Finally, we derive the final representation matrix Z by aggregating each modality-specific representation according to attention score β:

$$z_i = \sum_{m=1}^{M} \beta_i^m \cdot y_i^m. \quad (10)$$

After model training, the latent representation $z_i \in R^{d_3}$ can be used in various downstream analyses, including clustering, visualization, and DEG detection.

Model training of SpatialGlue

The resulting model is trained jointly with two different loss functions, i.e., reconstruction loss and correspondence loss. Each loss function is described as follows.

Reconstruction loss To enforce the learned latent representation to preserve the expression profiles from different modalities, we design an individual decoder for each modality to reverse the integrated representation Z back into the normalized expression space. Specifically, by taking output Z from the Between-Modality attention aggregation layer as input, the reconstructed representations $\widehat{H}_1^l$ and $\widehat{H}_2^l$ from the decoder at the l-th ($l \in \{1,2,\dots,L-1,L\}$) layer are formulated as follows:

$$\widehat{H}_1^l = \sigma(\tilde{A}_s Z_1^{l-1} W_{d1}^{l-1} + b_{d1}^{l-1}), \quad (11)$$

$$\widehat{H}_2^l = \sigma(\tilde{A}_s Z_1^{l-1} W_{d2}^{l-1} + b_{d2}^{l-1}), \quad (12)$$

where W_{d1}, W_{d2}, b_{d1}, and b_{d2} are trainable weight matrices and bias vectors, respectively. $\widehat{H}_1^l$ and $\widehat{H}_2^l$ represent the reconstructed expression matrices for the omics modalities 1 and 2, respectively.

SpatialGlue's objective function to minimize the expression reconstruction loss is as follows:

$$\mathcal{L}_{recon} = \gamma_1 \sum_{i=1}^{N} \left\| x_i^1 - \hat{h}_i^1 \right\|_F^2 + \gamma_2 \sum_{i=1}^{N} \left\| x_i^2 - \hat{h}_i^2 \right\|_F^2, \quad (13)$$

where x^1 and x^2 represent the original features of the modalities 1 and 2, respectively. γ_1 and γ_2 are weight factors that are utilized to balance the contribution of different modalities. Due to the differences of sequencing technologies and molecular types, the feature distributions of different omics assays can vary significantly. As such, the weight factors also vary between different spatial multi-omics technologies but are fixed for datasets obtained using the same omics technology.

Correspondence loss While reconstruction loss can enforce the learned latent representation to simultaneously capture the expression information of different modality data, it does not guarantee that the representation manifolds are fully aligned across modalities. To deal with the issue, we add a correspondence loss function. Correspondence loss aims to force consistency between a modality-specific representation Y and its corresponding representation $\hat{Y}$ obtained through the decoder-encoder of another modality. Mathematically, the correspondence loss is defined as follows:

$$\mathcal{L}_{corr} = \gamma_3 \sum_{i=1}^{N} \left\| y_i^1 - \hat{y}_i^1 \right\|_F^2 + \gamma_4 \sum_{i=1}^{N} \left\| y_i^2 - \hat{y}_i^2 \right\|_F^2, \quad (14)$$

$$\hat{Y}_1^l = \sigma\left(\tilde{A}_s\left(\sigma\left(\tilde{A}_s Y_1^{l-1} W_{d2}^{l-1} + b_{d2}^{l-1}\right)\right) W_{e2}^{l-1} + b_{e2}^{l-1}\right), \quad (15)$$

$$\hat{Y}_2^l = \sigma\left(\tilde{A}_s\left(\sigma\left(\tilde{A}_s Y_2^{l-1} W_{d1}^{l-1} + b_{d1}^{l-1}\right)\right) W_{e1}^{l-1} + b_{e1}^{l-1}\right). \quad (16)$$

where γ_3 and γ_4 are hyper-parameters, controlling the influences of different modality data. We set $\hat{Y}_1^0 = Y_1$ and $\hat{Y}_2^0 = Y_2$. $\sigma(\cdot)$ which is a nonlinear activation function, i.e., ReLU (Rectified Linear Unit).

Therefore, the overall loss function used for model training is defined as:

$\mathcal{L}_{total} = \mathcal{L}_{recon} + \mathcal{L}_{corr}$. (17)

Implementation details of SpatialGlue

For all datasets, a learning rate of 0.0001 was used. To account for differences in feature distribution across the datasets, a tailored group of weight factors $[\gamma_1, \gamma_2, \gamma_3, \gamma_4]$ was empirically assigned to each one. The weight factors were [1, 5, 1, 1] for the SPOTS mouse spleen dataset, [1,5,1,10] for the 10x Genomics Visium human lymph node dataset, [1, 10, 1, 10] for the Stereo-CITE-seq mouse thymus dataset, [1, 5, 1, 1] for the spatial-epigenome-transcriptome mouse brain dataset. We also provided a default parameter set that would work for most users on most data types. The training epochs used for the SPOTS mouse spleen, 10x Genomics Visium human lymph node, Stereo-CITE-seq mouse thymus, and spatial-epigenome-transcriptome mouse brain datasets were 600, 200, 1500, and 1600, respectively.

Data and detailed methods

For details on the datasets, downstream analyses, competing methods, and metrics employed, please see the supplementary file.

Data availability

The SPOTS mouse spleen data was obtained from the GEO repository (accession no. GSE198353, https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE198353) [4], the Stereo-CITE-seq mouse thymus data from BGI and the spatial-epigenome-transcriptome mouse brain data from AtlasXplore (https://web.atlasxomics.com/visualization/Fan) [3]. The details of all datasets used are available in the Methods section. The data used as input to the methods tested in this study, inclusive of the Stereo-CITE-seq and the in-house human lymph node data have been uploaded to Zenodo and is freely available at https://zenodo.org/record/7879713#.ZE3aOnZByUk.

Code availability

An open-source Python implementation of the SpatialGlue toolkit is accessible at https://github.com/JinmiaoChenLab/SpatialGlue.

Author contributions

J.C. conceptualized and supervised the project. Y.L. designed the model. Y.L. developed the SpatialGlue software. Y.L., K.S.A., and J.C. wrote the manuscript. Y.L., J.C., R. F., D. Z., R.S, S.C.Y., C.Z, H.X., and K.S.A. performed the data analysis. C.Z. and R.S. ran the Seurat WNN algorithm. Y.L. prepared the figures. J.C., N.R.J.G, L.G.N., and N.H. annotated and interpreted

the mouse thymus dataset. L.V.O and I.K. annotated the human lymph node datasets. D.G. generated the human lymph node dataset. S.L., Y.H., M.J., A.C., and X.X generated the mouse thymus dataset.

Acknowledgements

We thank Yingrou Tan for her assistance in mouse thymus data interpretation and Min Wu for his comments on the model. The research was supported by A*STAR under its BMRC Central Research Fund (CRF, UIBR) Award; AI, Analytics and Informatics (AI3) Horizontal Technology Programme Office (HTPO) seed grant (Spatial transcriptomics ST in conjunction with graph neural networks for cell–cell interaction #C211118015) from A*STAR, Singapore; Open Fund Individual Research Grant (Mapping hematopoietic lineages of healthy and high-risk acute myeloid leukemia patients with FLT3-ITD mutations using single-cell omics #OFIRG18nov-0103) from Ministry of Health, Singapore; National Research Foundation (NRF), Award no. NRF-CRP26-2021-0001; the National Research Foundation, Singapore, and Singapore Ministry of Health's National Medical Research Council under its Open Fund-Large Collaborative Grant ("OF-LCG") (MOH-OFLCG18May-0003). Singapore National Medical Research Council (#NMRC/OFLCG/003/2018).

Ethics declarations

The authors declare that there are no competing interests.

References

1. Liu, Y. *et al.* High-Spatial-Resolution Multi-Omics Sequencing via Deterministic Barcoding in Tissue. *Cell* **183**, 1665-1681.e18 (2020).

2. Liu, Y. *et al.* High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01676-0.

3. Zhang, D. *et al.* Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113–122 (2023).

4. Ben-Chetrit, N. *et al.* Integration of whole transcriptome spatial profiling with protein markers. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-022-01536-3.

5. Vickovic, S. *et al.* SM-Omics is an automated platform for high-throughput spatial multi-omics. *Nat. Commun.* **13**, 795 (2022).

6. Liao, S. *et al.* Integrated Spatial Transcriptomic and Proteomic Analysis of Fresh Frozen Tissue Based on Stereo-seq. *bioRxiv* 2023.04.28.538364 (2023) doi:10.1101/2023.04.28.538364.

7. Hudson, W. H. & Sudmeier, L. J. Localization of T cell clonotypes using the Visium spatial transcriptomics platform. *STAR Protoc.* **3**, 101391 (2022).

8. Janesick, A. *et al.* High resolution mapping of the breast cancer tumor microenvironment using integrated single cell, spatial and in situ analysis of FFPE tissue. *bioRxiv* 2022.10.06.510405 (2022) doi:10.1101/2022.10.06.510405.

9. Takei, Y. *et al.* Integrated spatial genomics reveals global architecture of single nuclei. *Nature* **590**, 344–350 (2021).

10. Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B. & Zhuang, X. Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell* **182**, 1641-1659.e26 (2020).

11. Liu, J. *et al.* Concordance of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *Life Sci. alliance* **6**, (2023).

12. He, S. *et al.* High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat. Biotechnol.* **40**, 1794–1806 (2022).

13. Dong, K. & Zhang, S. Deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder. *Nat. Commun.* **13**, 1739 (2022).

14. Hu, J. *et al.* SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).

15. Long, Y. *et al.* Spatially informed clustering, integration, and deconvolution of spatial transcriptomics with GraphST. *Nat. Commun.* **14**, 1155 (2023).

16. Hao, Y. *et al.* Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573-3587.e29 (2021).

17. Argelaguet, R. *et al.* MOFA+: A statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.* (2020) doi:10.1186/s13059-020-02015-1.

18. Gayoso, A. *et al.* Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).

19. Ashuach, T. *et al.* MultiVI: deep generative model for the integration of multimodal data. *Nat. Methods* **20**, 1222–1231 (2023).

20. Minoura, K., Abe, K., Nam, H., Nishikawa, H. & Shimamura, T. A mixture-of-experts deep generative model for integrated analysis of single-cell multiomics data. *Cell reports methods* **1**, 100071 (2021).

21. Ghazanfar, S., Guibentif, C. & Marioni, J. C. Stabilized mosaic single-cell data integration using unshared features. *Nat. Biotechnol.* (2023) doi:10.1038/s41587-023-01766-z.

22. Alexandre, Y. O. & Mueller, S. N. Splenic stromal niches in homeostasis and immunity. *Nat. Rev. Immunol.* (2023) doi:10.1038/s41577-023-00857-x.

23. Borges da Silva, H. *et al.* Splenic Macrophage Subsets and Their Function during Blood-Borne Infections. *Front. Immunol.* **6**, 480 (2015).

24. Backer, R. *et al.* Effective collaboration between marginal metallophilic macrophages and CD8+ dendritic cells in the generation of cytotoxic T cells. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 216–221 (2010).

25. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).

26. Kipf, T. N. & Welling, M. Semi-Supervised Classification with Graph Convolutional Networks. (2016) doi:10.48550/ARXIV.1609.02907.

Figure legends

Figure 1: Interpretable deep dual-attention model that enables accurate identification of spatial domains in simulated and real data. (a) Overview of the SpatialGlue framework. Spatial multi-omics experiment to simultaneously measure two distinct types of molecules, such as RNA and surface protein, while preserving spatial context of the tissue. SpatialGlue first uses the K-nearest neighbor (KNN) algorithm to construct a spatial neighbor graph using the spatial coordinates and a feature neighbor graph with the normalized expression data for each omics modality. Then for each modality, a GNN-encoder takes in the normalized expressions and the neighbor graph to learn two graph-specific representations by iteratively aggregating representations of neighbors. To capture the importance of different graphs, we designed a Within-Modality attention aggregation layer to adaptively integrate graph-specific representations and obtain a modality-specific representation. Finally, to preserve the importance of different modalities, SpatialGlue uses a Between-Modality attention aggregation layer to adaptively integrate modality-specific representations and output the final integrated representation of spots. (b) Spatial plots of the simulated data, from left to right: ground truth, generated raw data of individual modalities, and clustering results by single-cell and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. 'backgr' means background. (c) Density distribution of the simulated data modalities. (d) Modality weights of different modalities, denoting their importance to the integrated output of SpatialGlue. (e) Quantitative evaluation of methods with six supervised metrics. (f) Boxplots of the six metrics with the scores from 5 simulated datasets. (g) Manual annotation of human lymph node sample A1. (h) Spatial plots of lymph node sample A1, clustering of individual RNA and protein modalities (left), clustering results (right) from single and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. Note that the colors of clusters do not directly correspond to the same captured structures across different methods. (i) Boxplots of Moran's I score of the eight methods. (j) Jaccard similarity scores of the eight methods. (k) Boxplots of six supervised metrics with scores of clustering results with the number of clusters ranging from 4 to 11.

Figure 2: SpatialGlue dissects spatial-epigenome-transcriptome mouse brain samples at higher resolution. (a) Annotated reference of the mouse brain coronal section from the Allen Mouse Brain Atlas. (b) Spatial plots of the RNA-seq and ATAC-seq data with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, MultiVI, MOFA+, scMM, StabMap, and SpatialGlue. The annotated labels correspond to SpatialGlue's results and the clustering colors do not necessarily capture the same structures across other methods. The full names of the abbreviation used are, ctx: cerebral cortex, cp: caudoputamen, vl: lateral ventricle, lpo: lateral preopic area, aca: anterior cingulate area, ls: lateral septal nucleus, aco: anterior commissure, olfactory limb, acb: nucleus accumbens, cc: corpus callosum. (c) Boxplots of Moran's *I* score of the six methods. (d) Comparison of Jaccard similarity scores of the six methods. (e) Modality weights of different modalities, denoting their importance to the integrated output of SpatialGlue. (f) Spatial plots of the RNA-seq and CUT&Tag-seq

(H3K27ac) data with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, MultiVI, MOFA+, scMM, StabMap, and SpatialGlue. The annotated labels correspond to SpatialGlue's results and the clustering colors do not necessarily capture the same structures across other methods. (g) Boxplots of Moran's *I* score of the six methods. (h) Comparison of Jaccard similarity scores of the six methods. (i) Modality weights of different modalities, denoting their importance to the integrated output of SpatialGlue. (j) Heatmap of differentially expressed genes for each cluster (k) Heatmap of differentially expressed peaks for each cluster. (l) Heatmap of peak-to-gene links.

Figure 3: SpatialGlue accurately integrates multi-modal data from the mouse thymus (RNA and protein acquired with Stereo-CITE-seq) and mouse spleen (RNA and protein acquired using SPOTS). (a) Spatial plots of RNA and protein data (mouse thymus acquired with Stereo-CITE-seq) with unimodal clustering (left), and comparison of clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. The annotated labels correspond to SpatialGlue's results and the clustering colors do not necessarily capture the same structures across other methods. (b) Boxplots of Moran's *I* score of the eight methods. (c) Comparison of Jaccard similarity scores of the eight methods. (d) Modality weights of different modalities, denoting their importance to the integrated output of SpatialGlue. (e) Histology image of the mouse spleen replicate 1 sample. (f) Spatial plots RNA and protein data (mouse spleen acquired using SPOTS) with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration method, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. The full names of the abbreviations RpMΦ, MMMΦ, and MZMΦ are red pulp macro, CD169+ MMM, CD209a+ MZM, respectively. (g) Boxplots of Moran's *I* score of the eight methods. (h) Comparison of Jaccard similarity scores of the eight methods. (i) Modality weights of different modalities, denoting their importance to the integrated output of SpatialGlue. (j) Heatmap of differentially expressed ADTs for each cluster. (k) Normalized ADT levels of key surface markers for T cells (CD3, CD4, CD8), B cells (IgD, B220, CD19) and RpMΦ (F4_80, CD68, CD163). (l) Violin plots of two marker genes in the MMMΦ, MZMΦ, and RpMΦ clusters.

Supplementary figures

Figure S1: (a) SpatialGlue's within-modality weights for the importance of spatial and feature graphs with simulation data 1. (b) Simulated data 2 ground truth, unimodal clustering of modalities, and integration results from Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (c) Density distribution of the simulated data modalities. (d) Quantitative comparison of the eight methods with six measurement metrics, homogeneity, mutual information, V measure score, adjusted mutual information (AMI), normalized mutual information (NMI), and adjusted rand index (ARI). (e) SpatialGlue's between-modality weights explaining the importance of each modality to each cluster. (f) Within-modality weights for the importance of spatial and feature graphs.

Figure S2: (a) Simulated data 3 ground truth, unimodal clustering of modalities, and integration results from Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (b) Density distribution of the simulated data modalities. (c) Quantitative comparison of the eight

methods with six measurement metrics, homogeneity, mutual information, V measure score, adjusted mutual information (AMI), normalized mutual information (NMI), and adjusted rand index (ARI). (d) SpatialGlue's between-modality weights explaining the importance of each modality to each cluster. (e) Within-modality weights for the importance of spatial and feature graphs.

Figure S3: (a) Simulated data 4 ground truth, unimodal clustering of modalities, and integration results from Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (b) Density distribution of the simulated data modalities. (c) Quantitative comparison of the eight methods with six measurement metrics, homogeneity, mutual information, V measure score, adjusted mutual information (AMI), normalized mutual information (NMI), and adjusted rand index (ARI). (d) SpatialGlue's between-modality weights explaining the importance of each modality to each cluster. (e) Within-modality weights for the importance of spatial and feature graphs. (f) Ground truth, unimodal clustering of modalities, and integration results from Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (g) Density distribution of the simulated data modalities. (h) Quantitative comparison of the eight methods with six measurement metrics, homogeneity, mutual information, V measure score, adjusted mutual information (AMI), normalized mutual information (NMI), and adjusted rand index (ARI). (i) SpatialGlue's between-modality weight explaining the importance of each modality to each cluster. (j) Within-modality weights for the importance of spatial and feature graphs.

Figure S4: Evaluation of SpatialGlue on simulated triple omics data. (a) ground truth and spatial plots of modalities 1, 2, 3, and SpatialGlue. (b) Density distribution of simulated data modalities. (c) SpatialGlue's between modality weights explaining the importance of each modality to each cluster. (d) Within-modality weights for the importance of spatial and feature graphs.

Figure S5: Ablation study to validate the contribution of each component to the performance of the SpatialGlue model. The ablation study was conducted using the simulated data 1. (a) Ground truth. (b) Spatial clustering of modalities 1 and 2. (c) Comparison of SpatialGlue and its variants, i.e., using concatenation (C) instead of attention (A) for intra-modality integration (SpatialGlue-CA), using concatenation instead of attention inter-modality for integration (SpatialGlue-AC), and using concatenation instead of attention for both intra- and inter-modality integration (SpatialGlue-CC). (d) Quantitative evaluation of SpatialGlue and variants (CA, AC, CC) with the six supervised metrics. (e) Clustering results of SpatialGlue and the non-spatial variant (SpatialGlue w/o spatial). (f) Quantitative comparison of SpatialGlue with 'SpatialGlue w/o spatial'. (g) Comparison of SpatialGlue and the variant of SpatialGlue without PCA (SpatialGlue-full). (h) Quantitative comparison of SpatialGlue with 'SpatialGlue-full'.

Figure S6: Comparison between SpatialGlue and single-modal methods on simulated and real (mouse brain P22 sample acquired using spatial-ATAC-RNA-seq) data. (a) Ground truth of the simulated data. (b) Comparison between SpatialGlue and single-modal methods, SpaGCN, STAGATE, and GraphST on the simulated data. (c) Quantitative evaluation using six supervised metrics. (d) Comparison between SpatialGlue and single-modal methods on the mouse brain P22 sample data. (e) Comparison of Moran's *I* score. (f) Comparison of Jaccard Similarity scores.

Figure S7: (a) Comparison of clustering results with different numbers of neighbors *k* to illustrate SpatialGlue's sensitivity to parameters. (b) Supervised metrics on the clustering results. (c) Comparison of clustering results with different number of PCs to illustrate SpatialGlue's sensitivity to input dimensionality. (d) Supervised metrics on the clustering results. (e) Comparison of clustering results with different numbers of GNN layers to illustrate SpatialGlue's sensitivity to input dimensionality. (f) Supervised metrics on the clustering results. (g) Time complexity of SpatialGlue and competing methods (Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap).

Figure S8: (a) SpatialGlue's between-modality weight explaining the importance of each modality to each cluster for the lymph node A1 sample. (b) Within-modality weights for the RNA and protein modalities explaining the contributions of the spatial and feature graphs to each cluster. (c) Quantitative evaluation of SpatialGlue and competing methods. (d) Ground truth for the lymph node D1 sample. (e) Spatial plots of RNA and protein data (left), and clustering results of single-cell and spatial multi-omics methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (f) Comparison of Moran's *I* score. (g) Comparison of Jaccard Similarity scores. (h) Quantitative evaluation with six supervised metrics. (i) Boxplots of six supervised metrics for clustering results with number of clusters ranging from 4 to 11. (j) Between-modality weights explaining the importance of each modality to each cluster. (k) Within-modality weights for the RNA and protein modalities explaining the contributions of the spatial and feature graphs to each cluster.

Figure S9: Heatmap of differentially expressed ADTs for each cluster for the human lymph node A1 (a) and D1 (b) samples.

Figure S10: ADT intensity plots of the lymph node A1 sample.

Figure S11: ADTs intensity plots of the lymph node D1 sample.

Figure S12: (a) Separate spatial plots of all clusters identified by SpatialGlue in the mouse brain P22 sample (spatial-ATAC-RNA-seq). (b) Separate spatial plots of all clusters identified by SpatialGlue in the mouse brain P22 sample (spatial-CUT&Tag-RNA-seq, H3K27ac). (c) Within-modality weights for the RNA and ATAC modalities explaining the importance of the spatial and feature graphs to each cluster. (d) Within-modality weights for the RNA and CUT&Tag (H3K27ac) modalities explaining the importance of the spatial and feature graphs to each cluster. (e) Modality weights of Seurat when applied to the spatial-ATAC-RNA-seq sample. (f) Modality weights of Seurat when applied to the spatial-CUT&Tag-RNA-seq (H3K27ac) sample.

Figure S13: (a) Intensity plots of marker genes in the mouse brain P22 sample (spatial-CUT&Tag-RNA-seq, H3K27ac). (b) Normalized gene activity scores from Zhang et al. (c) Peak-to-gene links plots.

Figure S14: (a) Results of the mouse brain P22 sample acquired with RNA-seq and CUT&Tag-seq (H3K4me3). Spatial plots of data modalities with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, MultiVI,

MOFA+, scMM, StabMap, and SpatialGlue. (b) Comparison of Moran's *I* score. (c) Comparison of Jaccard Similarity scores. (d) Between-modality weights explaining the importance of each modality to each cluster. (e) Within-modality weights explaining the contributions of the spatial and feature graphs to each cluster for each modality. (f) Comparison of spatial clustering using Seurat with 10 and 50 PC dimensions in the mouse brain P22 spatial-ATAC-RNA-seq sample. (g) Comparison of SpatialGlue and its variants, i.e., SpatialGlue without reconstruction loss ('SpatialGlue w/o recon') and SpatialGlue without correspondence loss ('SpatialGlue w/o corr'), in the mouse brain P22 spatial-ATAC-RNA-seq sample. (h) Comparison of Moran's I score of SpatialGlue and its two variants.

Figure S15: Results of the mouse brain P22 sample acquired with RNA-seq and CUT&Tag-seq (H3K27me3). Spatial plots of data modalities with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, MultiVI, MOFA+, scMM, StabMap, and SpatialGlue. (b) Comparison of Moran's *I* score. (c) Comparison of Jaccard Similarity scores. (d) Between-modality weights explaining the importance of each modality to each cluster. (e) Within-modality weights explaining the contributions of the spatial and feature graphs to each cluster for each modality.

Figure S16: Additional results for the mouse thymus 1 sample. (a) dsDNA image. (b) Total mRNA counts. (c) Modality weight from Seurat when applied to the sample. (d) Within-modality weights of SpatialGlue explaining the contributions of the spatial and feature graphs to each cluster for each modality. (e) Separate spatial plots of all clusters identified by SpatialGlue. (f) Expression of marker genes and proteins for each cell type.

Figure S17: Results for the mouse thymus 2 sample. (a) Spatial plots of data modalities with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (b) Comparison of Moran's *I* score. (c) Comparison of Jaccard Similarity scores. (d) Between-modality weight explaining the importance of each modality to each cluster. (e) Within-modality weights explaining the contributions of the spatial and feature graphs to each cluster for each modality.

Figure S18: Results for the mouse thymus 3 sample. (a) Spatial plots of data modalities with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (b) Comparison of Moran's *I* score. (c) Comparison of Jaccard Similarity scores. (d) Between-modality weight explaining the importance of each modality to each cluster. (e) Within-modality weights explaining the contributions of the spatial and feature graphs to each cluster for each modality.

Figure S19: Results for the mouse thymus 4 sample. (a) Spatial plots of data modalities with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, StabMap, and SpatialGlue. (b) Comparison of Moran's *I* score. (c) Comparison of Jaccard Similarity scores. (d) Between-modality weight explaining the importance of each modality to each cluster. (e) Within-modality

weights explaining the contributions of the spatial and feature graphs to each cluster for each modality.

Figure S20: Heatmap of differentially expressed ADTs for each cluster from the mouse thymus 1 (a), 2 (b), 3 (c), and 4 (d) samples.

Figure S21: Intensity plots of ADTs for the mouse thymus 1 sample.

Figure S22: Intensity plots of ADTs for the mouse thymus 2 sample.

Figure S23: Intensity plots of ADTs for the mouse thymus 3 (a) and 4 (b) samples.

Figure S24: Results for the mouse spleen replicate 1 sample. (a) Spatial plots of SpatialGlue's clusters together (left) and separate (right). (b) UMAP plots of the RNA and protein data modalities (left), and spatial plot of SpatialGlue's clusters (right). (c) Comparison of fraction of nearest neighbors metric for each annotated cluster calculated by the different modalities (original RNA and protein expression). (d) Neighborhood enrichment of cell type pairs. (e) Cluster co-occurrence scores for each cluster at increasing distances. (f) Spatial plots of the RNA and protein data modalities. (g) Cross tabulation heatmap of the clustering labels between the RNA and protein data.

Figure S25: Results for the mouse spleen replicate 1(a-c) and 2(d-h) samples. (a) Cross tabulation heatmap for the number of clusters between the RNA and protein data. (b) Modality weights from Seurat. (c) Within-modality weights of SpatialGlue explaining the contributions of the spatial and feature graphs to each cluster for each modality. (d) Spatial plots of data modalities with unimodal clustering (left), and clustering results (right) from single-cell and spatial multi-omics integration methods, Seurat, totalVI, MultiVI, MOFA+, MEFISTO, scMM, StabMap, and SpatialGlue. (e) Comparison of Moran's *I* score. (f) Comparison of Jaccard Similarity scores. (g) Between-modality weight explaining the importance of each modality to each cluster. (h) Within-modality weights explaining the contributions of the spatial and feature graphs to each cluster for each modality.