OXFORD

*Briefings in Bioinformatics*, 2025, **26(3)**, bbaf210
https://doi.org/10.1093/bib/bbaf210
**Case Study**

# StereoMM: a graph fusion model for integrating spatial transcriptomic data and pathological images

Bingying Luo[1,2,‡], Fei Teng[1,2,‡], Guo Tang[1], Weixuan Cen[2], Xing Liu[1,2], Jinmiao Chen[3], Chi Qu[1], Xuanzhu Liu[1,2], Xin Liu[1], Wenyan Jiang[1], Huaqiang Huang[1], Yu Feng[4,5], Xue Zhang[1], Min Jian[2], Mei Li[2], Feng Xi[1], Guibo Li[1,2], Sha Liao[2,§], Ao Chen[1,2,§], Weimiao Yu[6,§], Xun Xu[1,2,7,§], Jiajun Zhang[1,2,§,*]

[1]BGI Research, Chongqing, No. 313, Jinyue road, Jiulongpo District, Chongqing 401329, China
[2]BGI Research, Shenzhen, No. 9, Yunhua Road, Yantian District, Shenzhen 518083, China
[3]Center for Computational Biology and Program in Cancer and Stem Cell Biology, Duke-NUS Medical School, 8 College Road, Singapore 169857, Singapore
[4]State Key Laboratory of Genome and Multi-omics Technologies, BGI Research, No. 9, Yunhua Road, Yantian District, Shenzhen 518083, China
[5]BGI Collaborative Center for Future Medicine, Shanxi Medical University, No. 1258, Xinjiannan Road, Yingze District, Taiyuan 030001, China
[6]School of Biological Science, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551, Singapore
[7]BGI Research, Hangzhou, No. 203, Zhenzhong Road, Xihu District, Hangzhou 310030, China

*Corresponding author. Building 11, Beishan Industrial Zone, Yantian District, Shenzhen 518083, China. E-mail: zhangjiajun1@genomics.cn
‡Bingying Luo and Fei Teng contributed equally to this work.
§Sha Liao, Ao Chen, Weimiao Yu, Xun Xu and Jiajun Zhang are corresponding senior authors.

## Abstract

Spatial omics technologies, generating high-throughput and multimodal data, have necessitated the development of advanced data integration methods to facilitate comprehensive biological and clinical treatment discoveries. Based on the cross-attention concept, we developed an AI learning based toolchain called StereoMM, a graph based fusion model that can incorporate omics data such as gene expression, histological images, and spatial location. StereoMM uses an attention module for omics data interaction and a graph autoencoder to integrate spatial positions and omics data in a self-supervised manner. Applying StereoMM across various cancer types and platforms has demonstrated its robust capability. StereoMM outperforms competitors in identifying spatial regions reflecting tumour progression and shows promise in classifying colorectal cancer patients into deficient mismatch repair and proficient mismatch repair groups. The comprehensive inter-modal integration and efficiency of StereoMM enable researchers to construct spatial views of integrated multimodal features efficiently, advancing thorough tissue and patient characterization.

**Keywords:** spatial omics; multimodal data; deep learning; attention mechanism; molecular characteristics; patient classification

## Introduction

In the processes of diagnosis, evaluation, and therapeutic strategy formulation, physicians synthesize data from multiple sources. These data encompass three key dimensions: molecular biological information (molecular level), medical imaging information from clinical exams (tissue level), and clinical information from medical practice (patient level) (Supplementary Fig. 1a). Spatial relationships are powerful tools in precision medicine. Researchers successfully used spatial analysis to predict the effectiveness of anti-PD-1 treatments in metastatic melanoma [1], which highlights the importance of spatial information in disease treatment. The development of spatial omics has further advanced this field by generating multimodal data that retains in situ information. This allows the alignment of various data types on spatial coordinates, providing a more detailed and accurate representation of the dynamic biological processes of high dimensions. Nevertheless, the practicality of spatial omics such as spatial transcriptomic (ST) data is constrained by limitations such as low total transcriptions per cell, significant data noise, necessitating the integration of additional modal data for a comprehensive analysis [2, 3]. New methodologies capable of integrating multiple modalities (MM) data can help overcome limitations of existing methods, enable the insight of biological governing mechanisms and support the development of novel treatment strategies. The advancements in multimodality data acquisition and AI technology have partially opened a door of exciting possibilities in the field of precision medicine. A prevalent trend is to combine MM data, using bioinformatics and AI algorithms. AI-assisted diagnostic models have been developed and implemented for various disease classifications [4–7].

The tumour microenvironment in solid tumour is highly complex, which is the one of the root causes of poor response rate and efficacy of Immunotherapy treatment. We are in need of comprehensive analysis of multimodal information to fully understand its intricacies. By integrating multiple aspects of patient information, AI algorithms can perform nonlinear analysis on these spatially aligned or unaligned complex datasets (Supplementary Fig. 1b), enabling more precise tumour heterogeneity assessments. This facilitates the discovery of novel

**Received:** December 9, 2024. **Revised:** March 27, 2025. **Accepted:** April 10, 2025

targets and biomarkers, expediting new drug development, and allows for accurate patient stratification in guiding medication decisions (Supplementary Fig. 1c). In alignment with this vision, our research endeavours extend to a granular level, and we are dedicated to applications such as tumour microenvironment analysis and exploration of spatial domains, aiming to uncover the complete landscape of the disease (Supplementary Fig. 1d).

Several algorithms have been designed to integrate information from MM of the ST data. For example, stLearn [8] is a widely used spatial transcriptomics analysis tool. However, it does not perform appropriate weighting when normalizing histological images with spatial location. spaGCN utilizes graph convolutional network (GCN) to model spatial relationships [9]. It has limited capabilities in feature extraction because it utilizes the pixel values of the three channels of the image and ignores the high-level features of morphology. STAGATE [10] utilizes a Graph Attention Network (GAT) for model construction, yet its reliance on reconstructing the gene expression matrix alone limits its ability to effectively capture useful multi-modal information. Software such as MUSE [11] and SEDR [12] employ architectures underpinned by autoencoders to learn a low-dimensional representation of multimodal data, but such integration relies entirely on non-linear activation functions. stMDA [13] employs Maximum Mean Discrepancy (MMD) to align the distributions of multi-modal data; however, MMD may struggle to capture complex, high-dimensional relationships, potentially limiting its ability to fully integrate multi-modal data. Current software assumes that spatially adjacent cells or tissues are similar but fails to consider the consistency of distant regions, potentially leading to an incomplete understanding of certain biological phenomena. For example, in the context of immune cell infiltration in tumours, immune cells located in different regions of the tumour microenvironment may exhibit similar activation states or functional properties. Ignoring these spatial distant relationships can result in missing critical insights into the overall immune response and its impact on tumour progression and treatment outcomes. Additionally, downstream analysis is limited to individual sample evaluations and lacks cross-sample comparisons, which are crucial for accurately understanding disease characteristics in real-world clinical applications.

To overcome these limitations, we designed the StereoMM method, which integrates RNA expression data, H&E image information, and tissue in situ locations in the spatial transcriptome via attention mechanisms and graph neural networks (GNN). This approach achieves comprehensive integration from molecular to tissue structure, obtaining multimodal joint embeddings to provide a more complete biological insight. Specifically, the attention mechanism dynamically allocates weights across the entire slice to highlight key features across different modalities in a manner that considers spatial distant relationships. This approach allows the model to capture and leverage long-range similarities, which can be crucial for identifying patterns and interactions that are not immediately adjacent but still significantly impact the overall analysis. By extracting these attention weights, StereoMM can provide quantitative standards for decision-making, significantly enhancing the interpretability of results for cancer researchers. The model captures interactions between different modality and provides a more accurate representation for downstream analysis. Furthermore, we have developed a sample-level classification method to better address clinical requirements.

StereoMM was tested using mouse brain tissue, demonstrating its capability to discern fine tissue architecture, while highlighting its advantage in computational speed. We substantiated the efficacy of StereoMM through conceptual validation across multiple cancer datasets from diverse platforms. Utilizing data from human lung adenosquamous carcinoma obtained using Stereo-seq and human breast cancer from 10X Visium, we showed the superior performance of StereoMM in spatial domain recognition over competing software, and its ability to reveal tumour heterogeneity. We also used StereoMM to accurately classify patients with colorectal cancer data, effectively differentiating between patients with deficient mismatch repair (dMMR) and proficient mismatch repair (pMMR). StereoMM exhibited exceptional performance in identifying spatial domains and showed potential for patient stratification, demonstrating its superiority over existing methodologies, and its potential for predictive biomarker discovery.

# Results

## Overview of the StereoMM framework

Spatial transcriptomics methodologies, particularly high-resolution Stereo-seq [14], have transformed our understanding of tissue heterogeneity. Integrating spatially aligned data into a multimodal fusion algorithm framework allows for the correlation of these data (gene expression and tissue imaging), enabling the discernment of functional changes based on structural differences within our visibility. StereoMM structure leverages a cross-attention mechanism to focus on long-range dependencies between modalities (Fig. 1a), rather than just spatial neighbours. StereoMM not only identifies slice-level spatial domains and prognosis-related features but also excels in patient-level classification (Fig. 1b), setting it apart from existing software. StereoMM is trained using a graph-based self-supervised approach. It generates a feature representation that combines multiple modalities including spatial transcriptomic data and H&E imaging data. Overall, StereoMM is composed of two primary modules as in Fig. 1c: I). a cross-attention module for inter-modal information interaction and; II). a variational graph autoencoder (VGAE) module for aggregating spatial neighbours' information and self-supervised training.

Specifically, the training process is divided into the following five steps: (i) For the transcriptome and H&E images, a unimodal feature extractor is employed to extract $s$-dimensional unimodal features, and generate two feature matrices ($X_t \in R^{n\times s}$ for transcriptome, and $X_m \in R^{n\times s}$ for morphology, where $n$ represents the number of bins or spots). (ii) These features are then fed into the attention module, where the information between modalities is integrated using the cross-attention mechanism as in Fig. 1d. This integration results in an $s$-dimensional output that enhances the interaction between modalities ($X_{ta} \in R^{n\times s}$ for transcriptome, and $X_{ma} \in R^{n\times s}$ for morphology). (iii) The feature matrices from both modalities are concatenated ($X = X_{ta} \oplus X_{ma}$) and used as input for the node features of the graph autoencoder. (iv) To incorporate spatial location information, a Spatial Neighbour Graph (SNG) is generated based on the physical distance. The SNG serves as the input for the adjacency matrix in the graph autoencoder. (v) The graph data is subjected to a VGAE module, as illustrated in Fig. 2e. The encoder utilizes GNN to learn the distribution of node vector representations. These representations are then sampled from the distribution, and the graph is reconstructed using the decoder. The output of the decoder is $\hat{X}$.

The optimization of StereoMM is guided by two functions that minimize data loss: a self-supervised reconstruction loss and a regularization loss that forces the distribution of latent representation space. The implementation of a self-supervised

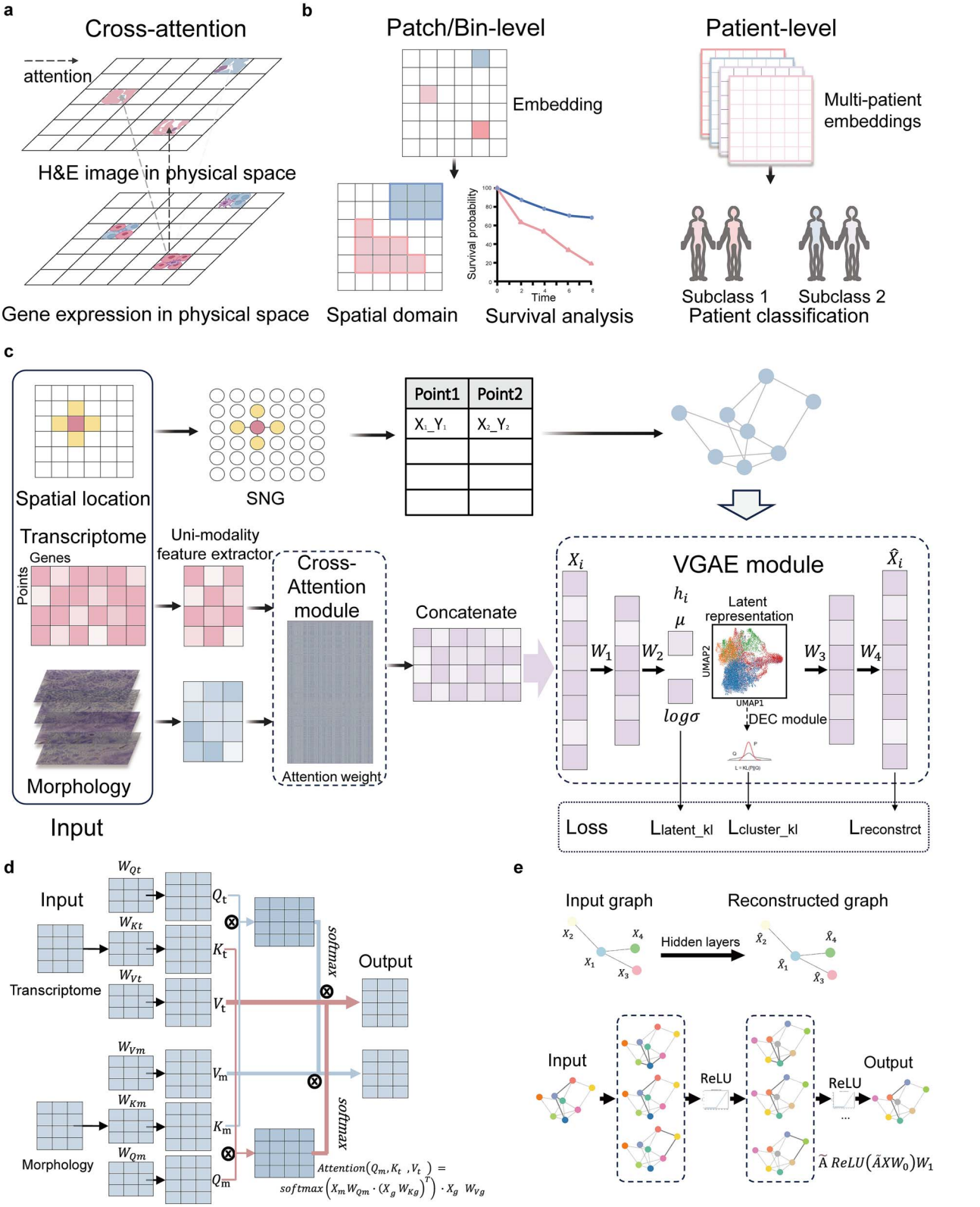


Figure 1. **Schematic overview of StereoMM**. (**a**) Simplified schematic of the cross-attention mechanism for inter-modal information interaction. The colour intensity of the arrows indicates the magnitude of the attention weights. This method can capture long-range dependencies between modalities. (**b**) Multimodal latent representations can be utilized for downstream analysis at both the patch/bin-level and the patient-level. (**c**) The overall framework of StereoMM. The three input modalities include spatial coordinates, gene expression matrix, and image patches. The attention module and VGAE module generate a low-dimensional latent representation which can be used for downstream tasks. (**d**) The formal representation of the cross-attention module in StereoMM. In this module, each individual modality generates its own set of queries (Q), keys (K), and values (V). The Q from one modality is used to query the K and V from another modality. **(e)** the VGAE module in StereoMM aggregates spatial information and each modality feature, and reduces the dimensionality of the original features through the encoder to obtain the final latent representation.

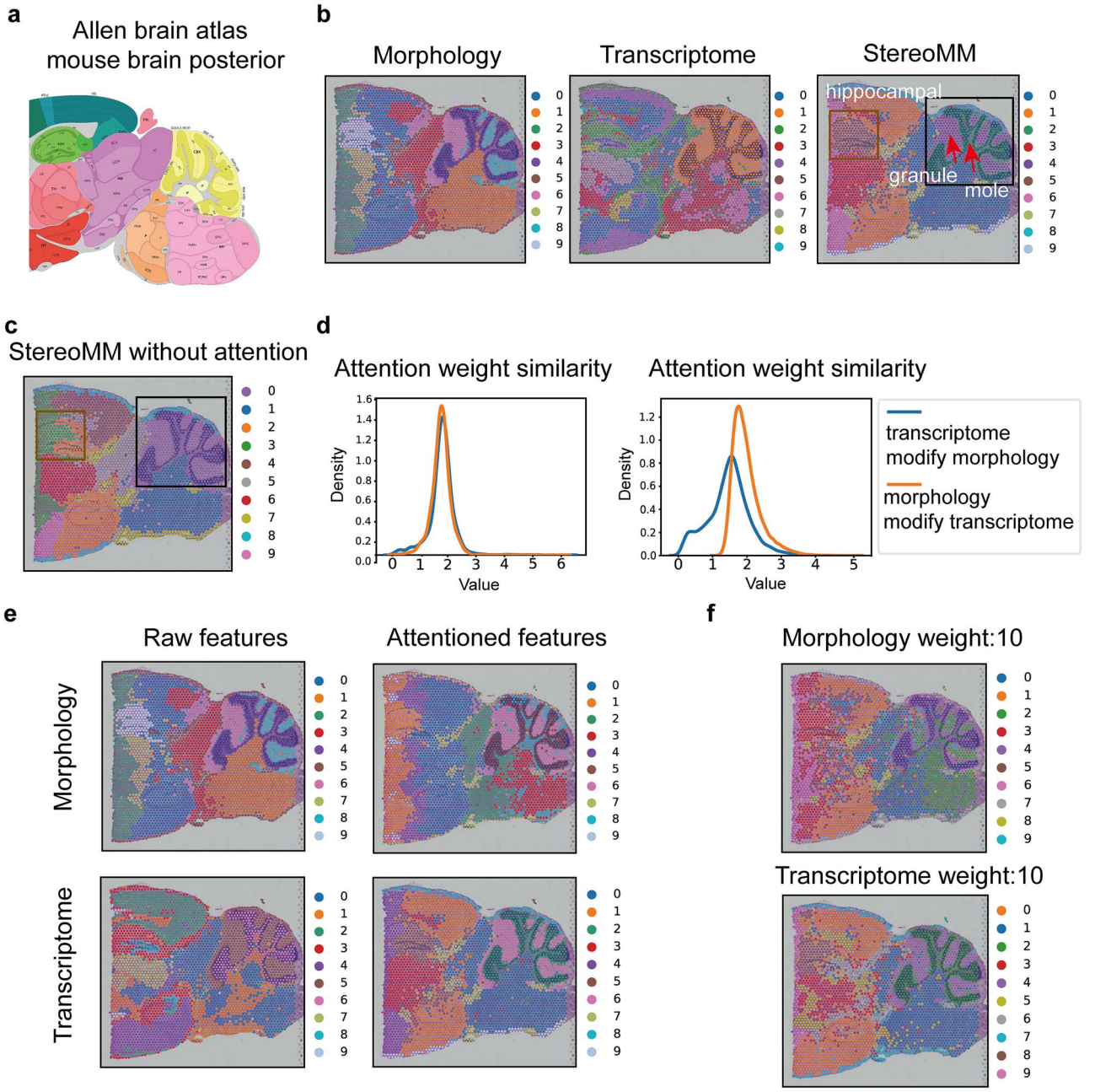


Figure 2. **Ablation experiment of attention module.** (**a**) The corresponding anatomical Allen mouse brain Atlas (https://atlas.brain-map.org/). (**b**) Spatial domains identified by StereoMM. The left box indicates the hippocampus region. The right box denotes the lobules area, arrows indicate the mole and granule regions respectively. (**c**) Spatial domains identified by StereoMM without attention module. (**d**) The correlation between attention-enhanced features and final latent representations Z. Left: The results on mouse brain slide. Right: The results on slide 3 of lung cancer. (**e**) The features before and after the attention module are used to identify the spatial domain. (**f**) The spatial domain recognized after manually setting the modality weight parameters.

algorithm is accomplished by the autoencoder in VGAE, with a reconstruction loss function that promotes a high degree of similarity between the generated output $\hat{X}$ and the original input matrix ($X$), ensuring that the output closely mirror the input. The autoencoder ensures that the latent features learned by the encoder function preserves the maximum information from the original multimodal input, and a decoder can reconstruct the original input through these latent features. The algorithm assesses regularization loss, also known as the Kullback–Leibler (KL) divergence, to promote compactness and smoothness in latent representations space, bolstering model generalization and sample continuity.

By extracting the latent representation from the VGAE module, a high-quality, low-dimensional representation ($Z \in R^{n \times d}$, where $d$ represents the feature dimension after dimensionality reduction) of the graph data is obtained. This feature representation $Z$ can be effectively utilized for various downstream analyses, including clustering, trajectory analysis, and more.

## Systematic model evaluation

We used a mouse brain tissue with intricate tissue structures as test sample for conducting an ablation experiment of attention module. Firstly, we demonstrated that StereoMM outperforms individual modalities alone. We used anatomical reference

annotations from the Allen Mouse Brain Atlas [15] as ground truth shown in Fig. 2a. StereoMM accurately identified the hippocampal structures and differentiated mole and granule areas in the lobules shown by the rectangle in Fig. 2b. None of the single modalities could independently identify this specific region.

We then performed ablation experiments on the mouse brain sample to demonstrate the effectiveness of the attention module. Without the interactive ability of the attention module, the mole and granule areas in the lobules could no longer be distinguished, identification of the hippocampal structures was also blurred, and more noise was introduced. Detailed comparisons are shown by the boxes in Fig. 2b and c. To clarify the role of the attention module in enhancing inter-modal information interaction, we extracted the weight matrix and computed its correlation with the final output (Z). This approach illuminates how our network assesses and assigns significance to individual modal features during data integration. In the mouse brain data, morphological similarity was on par with transcriptomic similarity, indicating that the model integrated the two modalities in a balanced manner. StereoMM was also tested on the lung cancer data from Stereo-seq, where the correlation between morphological features and the latent features is higher, suggesting that the model assigned a higher weight to the morphological features (Fig. 2d). In order to further illustrate the capabilities of the attention module, we extracted the features generated by the attention module for visualization shown in Fig. 2e. The attention module enhances the alignment of unimodal features, as reflected by the total cosine similarity's increase from −26.76 to −13.91, signifying improved information exchange between modalities.

We also provided a hyperparameter that weights the contribution of individual modalities based on their relative importance, thereby enhancing the model's guidance with prior knowledge. By setting custom weights for transcriptomic features, we maintain the flexibility of the model during the fusion process. As transcriptomic weight increased, the final output of the model became more similar to the transcriptome in Fig. 2f (ARI, from 0.17 to 0.29).

To quantitatively evaluate the performance, we utilized 12 human dorsolateral prefrontal cortex (DLPFC) datasets as a standard reference. These datasets include detailed manual annotations of distinct tissue regions, which we treated as the ground truth. We compared the spatial domain identification performance of StereoMM against STAGATE, stMDA, spaGCN, stLearn, MUSE, SEDR, and individual transcriptomic and morphological (Supplementary Fig. 2a). StereoMM achieved the highest ARI and NMI, demonstrating its superior accuracy in spatial domain identification (Supplementary Fig. 2b-c).

The attention module showcased distinct benefits in terms of enhancing model performance and interpretability. Notably, it provided a clear method for adjusting weights for individual modalities. The architecture based on attention and GNN used by our structure helped capture and combine information that could not be obtained from either mode alone. Quantitative evaluation on 12 DLPFC datasets demonstrated superior performance by StereoMM, achieving the highest adjust Rand Index (ARI) and normalized mutual information (NMI) compared to existing methods and individual modality.

## StereoMM improves the performance of domain identification in stereo-seq data of human lung adenosquamous carcinoma

To evaluate the accuracy of tissue identification and perform quantitative assessment of StereoMM, we conducted an analysis using lung adenosquamous carcinoma data generated from the Stereo-seq platform [16]. The data was meticulously annotated by pathologists into three distinct sections: lung adenocarcinoma (LUAD), lung squamous cell carcinoma (LUSC), and mixed areas as shown in Fig. 3a. To reduce the computational burden, we divided the data into four slices (Supplementary Fig. 3a).

We first tested the impact of different hyperparameters on the accuracy of the results. StereoMM provided three types of GNN, including GCN, GAT, and graph sample and aggregate (GraphSAGE). Although there was no statistical difference, GCN achieved the optimal results in ARI and NMI as in Fig. 3b, indicating the highest consistency with ground truth (Supplementary Fig. 3b). Users can define hidden layers and customize the number of nodes in each layer in the VGAE module. We assessed eight configurations: first layer (2048 or 1024 nodes), second layer (256 or 128 nodes), and third layer (50 or 100 nodes) (Fig. 3c, Supplementary Fig. 3c). In general, the performance of StereoMM was robust under the choice of different number of nodes (ARI and NMI, Anova, $P = 1$).

We also performed another benchmark analysis to compare the performance of the six competing algorithms (Supplementary Fig. 4a). By normalizing all algorithm results to same clusters, it was found that StereoMM significantly enhances unimodal analysis accuracy and reduces data noise (Fig. 3d). UMAP visualizations showed StereoMM clear separation of annotated categories, unlike the mixed state of the original transcriptome (Fig. 3e, Supplementary Fig. 5a). StereoMM outperformed previous spatial clustering methods in spatial recognition, achieving the highest ARI ($0.32 \pm 0.07$) and NMI ($0.34 \pm 0.05$) scores, indicating the highest consistency with manual annotations (Fig. 3f). Cluster internal evaluation metrics confirmed its reliable performance (Supplementary Fig. 5b). We compared the execution times of various software tools (Supplementary Fig. 5c) and found StereoMM to be the fastest. Using simulated data, StereoMM consistently exhibited the shortest runtime, regardless of cell count (Supplementary Fig. 5d), highlighting its superior processing speed.

To evaluate clinical utility of StereoMM, we compared its precision in delineating clinically significant regions with single transcriptomics, focusing on tumour subtype-specific markers. Spatial domains recognized by StereoMM showed higher accuracy for LUAD markers ($P = 0.0053$) although not for LUSC markers [17] ($P = 0.53$) (Supplementary Fig. 6a-e). Using WGCNA, we identified key modules and genes, such as CLDN3, CLDN4, and KRT19, linked to poor LUAD prognosis. For LUSC, hub genes S100A7, S100A8, and S100A9 were identified, highlighting their potential significance (Supplementary Fig. 7-8).

In summary, the architecture of StereoMM model demonstrated resilience, efficiency, and accuracy under various parameter configurations. A fair comparison of results showed that the recognition ability in the spatial domain of StereoMM was significantly better than that of a single modality or any competing software. Simultaneously, StereoMM can assist in identifying significant genes and putative targets related to the initiation and progression of tumours.

## StereoMM dissects breast cancer heterogeneity and identifies potential prognostic factors

To assess the compatibility of StereoMM with various spatial transcriptomics platforms, we applied it to an open-access dataset generated from fresh frozen invasive ductal carcinoma breast tissue using the 10x Visium spatial platform. Compared to transcriptome and morphology only, StereoMM not only revealed the

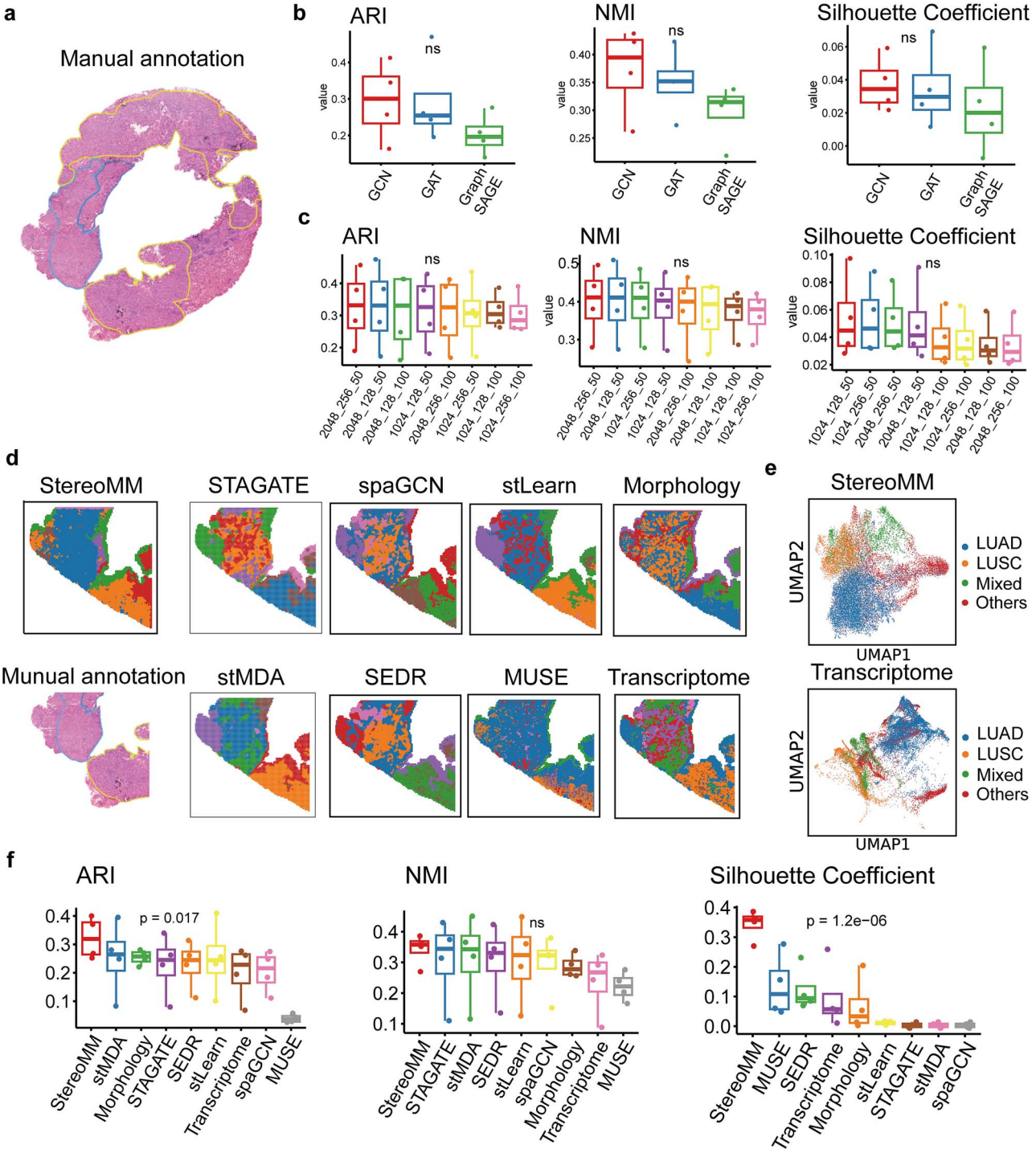


Figure 3. **StereoMM improves recognition performance of human lung adenocarcinoma pathological regions.** (**a**) Manual annotation by pathologist. Area circled by red marker showed AC phenotype, blue displayed SCC phenotype. Area enclosed in green presented mixed AC and SCC phenotypes. (**b**) Boxplots of ARI, NMI and LISI values for three GNN types, each evaluated on 4 lung cancer slides. The center line, box lines, and whiskers of the boxplot represent the median, upper and lower quartiles, and 1.5× interquartile range, respectively. (**c**) Boxplots of ARI, NMI and LISI values for different numbers of nodes per layer, each type evaluated on 4 lung cancer slides. (**d**) Manual annotation and the spatial domain identified by all algorithms on slide 4. (**e**) UMAP visualizations of transcriptome and latent representations generated by StereoMM. (**f**) Boxplot of LISI, ARI, NMI, and silhouette coefficient scores for seven methods in all 4 lung cancer slides. The center line, box lines, and whiskers of the boxplot represent the median, upper and lower quartiles, and 1.5× interquartile range, respectively.

clear clustering structure which was consistent with the manual annotation (Fig. 4a-b), but also specifically identified tumour boundary area as a separate domain (domain 2 in Fig. 5b).

To further investigate the intricate tumour microenvironment within spatial compartments identified by StereoMM, we performed a correlation analysis of clusters, and discovered two tumoural regions (Fig. 4c-d). Two regions named as tumour 1 and tumour 2, tumour 1 included domain 1 and 4 and tumour 2 included domain 0,3,9 and 10. We characterized the copy number variation and stemness in two tumoural regions. Tumour 1 displayed a distinctively higher inferCNV score (Fig. 4e, Student's t-test, $P = 6.44e\text{-}12$) and a higher score of CytoTRACE (Fig. 4f, Student's t-test, $P = 5.69e\text{-}236$), indicating tumour heterogeneity in accumulation of mutation and stemness. Cancer stemness has been reported to be affected by EMT states [18]. Tumour cells were annotated by cell2location (Supplementary Fig. 9a), and defined distinct EMT cell states ranging from epithelial (E-cad + VIM-), hybrid EMT (E-cad + VIM+) and mesenchymal (E-cad- VIM+) [19, 20] (Fig. 4g). We observed an increased proportion of the hybrid EMT cell types and decreased the proportion of epithelial cell types in tumour 1, indicating the high risk of infiltration and metastasis. Investigated the expression of EMT markers to prove EMT characteristics (Fig. 4h).

According to published information, the patient was diagnosed as ER + PR-HER2+ breast cancer. Therefore, we further investigated the differentially expressed genes (DEGs) ($|\log \text{fold change}| \geq 0.25$; $P < 0.05$) between tumour 1 and 2 (Fig. 4i). With GSEA analysis, tumour 1 region exhibited downregulation of 'INTERFERON GAMMA RESPONSE', 'ESTROGEN RESPONSE LATE' and 'ESTROGEN RESPONSE EARLY' (Fig. 4j). These pathways can interact with each other and are associated with the prognosis and treatment response of breast cancer [21, 22]. Meanwhile, we observed significant downregulation of SEMA3B and TFF1 in tumour1 (Fig. 4k-l), indicative of tumour suppressor functions in tumour 1. We also validated the prognostic value of SEMA3B and TFF1 using survival data from TCGA cohort of 333 HER2+ BC patients (Fig. 4m), suggesting the higher value of SEMA3B and TFF1 was associated with better prognosis, which was aligned with previous publications [23, 24]. Additionally, we identified a domain resembling tertiary lymphoid structure (Supplementary Fig. 9b-d), consistent with findings from other studies [25].

In summary, analysis of StereoMM clusters revealed regional and biological differences reflecting tumour progression and raised the hypothesis that heterogeneity in the stemness and EMT states, which might be with high risk of metastasis and resistance to therapy across histologic subtypes.

## StereoMM identifies consistent regions that are overlooked by pathological annotations

The applicability of StereoMM across a broad spectrum of cancer types has been further substantiated using colorectal cancer (CRC) data [26]. In CRC patient sample P19, domains identified by individual transcriptomes and StereoMM were annotated according to actual pathological annotations. The annotations by StereoMM domains demonstrated higher consistency with the pathological annotations (Fig. 5a-b, Supplementary Fig. 10a, ARI: 0.19 versus 0.28).

Intriguingly, StereoMM observed some distinct tumour regions extending into stromal areas. Histopathological examination with H&E staining revealed that these regions had cellular morphologies different from the stroma (Fig. 5c). To elucidate the properties of these tumour regions (Fig. 5d), DEGs were compared between these distinct tumour regions and both stromal and other tumour regions. These tumour regions exhibited far more upregulated genes compared to downregulated genes when compared with stroma. Conversely, this expression pattern was inverted when compared with other tumour domains (Fig. 5e, Supplementary Fig. 10b-c). Subsequent pathway enrichment analysis indicated that these areas upregulated classic tumour pathways such as 'MYC TARGET' when compared to the stroma. In comparison with other tumour areas, it exhibited discordant trends across pathways, indicating the heterogeneity within a single-sample tumour landscape (Fig. 5f). Specifically, these tumour regions upregulate genes associated with invasion and downregulate inhibitory immune-related genes (Supplementary Fig. 11a-b). An analysis of cell communication within a 200-micrometer perimeter of these regions (Fig. 5g) revealed interactions involving the IGF2-IGF2R and COL1A1-ITGA gene families (Fig. 5h), which have been previously documented to promote the progression of CRC [27, 28].

A comparable phenomenon was observed in CRC patient sample P36. StereoMM identified a stromal region surrounded by tumour areas (Fig. 5i-j), with distinct morphological features in H&E images (Fig. 5k). Differential gene expression analysis revealed minimal variation between this stromal region and others, with only 8 DEGs (Fig. 5m). Pathway enrichment analysis indicated significant downregulation of malignancy-associated processes, such as 'ANGIOGENESIS' and 'MYC TARGET', in the stromal region compared to the tumour areas (Fig. 5n). This may be due to its unique location bordering the tumour areas.

In summary, the annotation and comparison of the tumour and stromal regions demonstrate the ability of StereoMM to identify and characterize previously unnoticed regions in CRC samples. The biological activity, pathway enrichment, and cellular communication analyses demonstrate StereoMM's ability to detect tumour regions within the stromal area, which may be missed by traditional methods. This highlights the potential of StereoMM as a valuable tool for comprehensive analysis and understanding of the tumour microenvironment.

## StereoMM enables accurate classification of patients into different subtypes

Patient stratification contributes significantly to precision medicine in regards of diagnosis, treatment decision and post-surgery care. We collected 4 CRC samples containing patient with information of molecular subtyping. We applied StereoMM to Stereo-seq data from mentioned cohort. The result showed StereoMM possess the ability to distinguish patients with different molecular typing by constructing featured spatial domains (Fig. 6a). To aggregate the microenvironment relationships at the tissue level into a holistic representation at the patient level, we proposed a patient-level graph representation with StereoMM domains as nodes and inter-domain correlations as edges (Fig. 6b). The similarity between each graph (patient) was calculated using graph kernel methods.

We applied various methods to cluster the graph kernel results, including aggregating multiple clustering methods and using consensus clustering for sample sampling (Fig. 6c). Results from different clustering approaches showed high robustness. StereoMM successfully separated two dMMR patients from two pMMR patients (Fig. 6d). Utilizing transcriptome or morphological features alone following the above process fails to achieve successful classification (Supplementary Fig. 12a-f). To deeply explore the featured domain shared by patients belonging to the same group, we introduce graph node matching to find similar StereMM domains between patients. Figure 6e showed the graph node matching results for patients P55 and P59. Next,

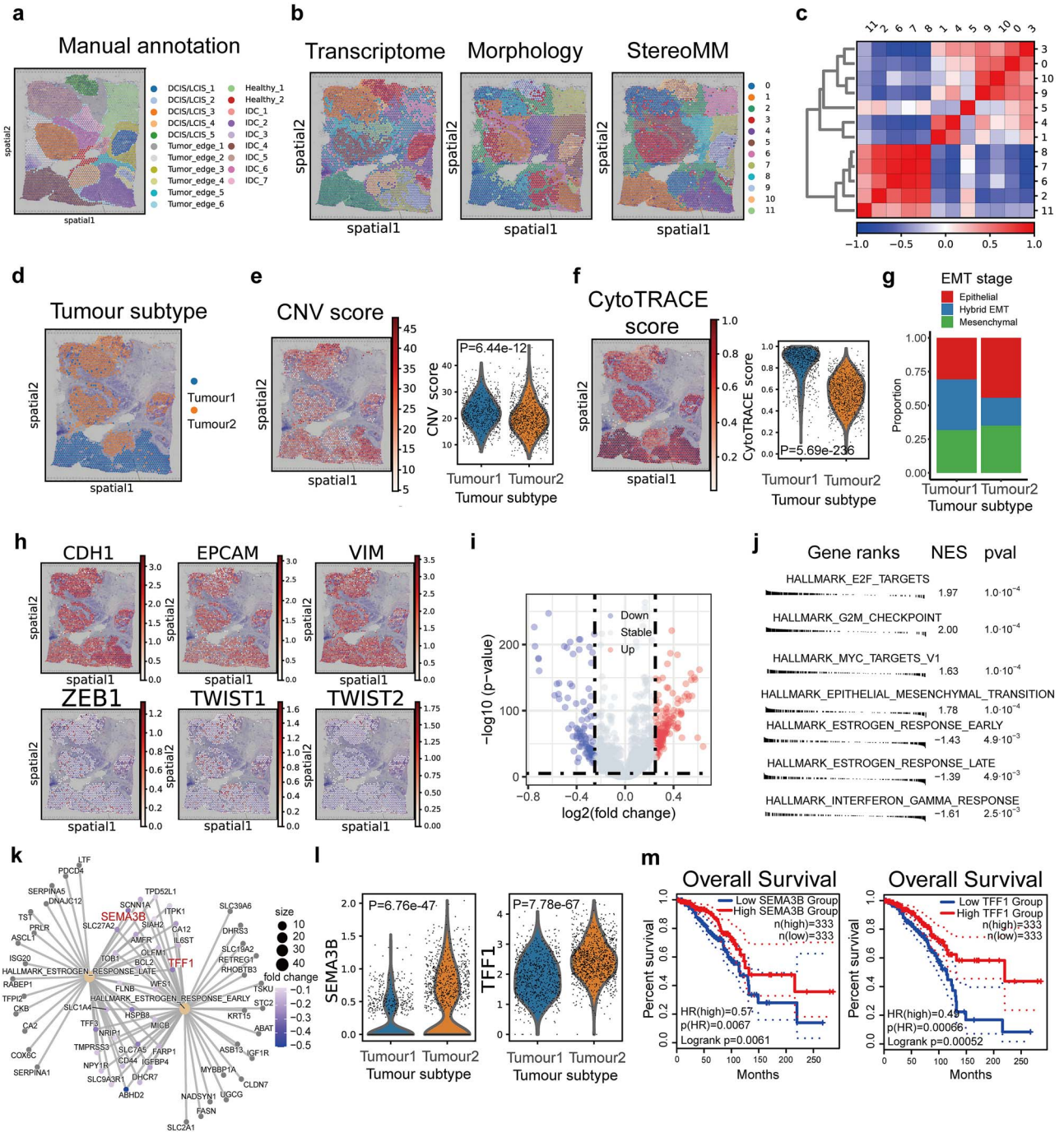


Figure 4. **StereoMM dissects breast cancer heterogeneity.** (**a**) Manual pathological annotation based on H&E staining of human breast cancer data. IDC, invasive ductal carcinoma; DCIS, ductal carcinoma in situ; LCIS, lobular carcinoma in situ; tumour edge; healthy region. (**b**) Spatial domains identified by each single modality (left: Transcriptome; middle: Morphology) and StereoMM (right). (**c**) Heatmap of Pearson correlation coefficients between domains. (**d**) Spatial location of tumour subtypes (tumour 1 and tumour 2). (**e**) CNV scores calculated by inferCNV for different tumour subtypes. (**f**) Stemness calculated by CytoTRACE for different tumour subtypes. (**g**) Proportion of EMT status in different tumour subtypes. (**h**) Spatial location of the expression of EMT-related marker genes. (**i**) volcano plot visualization of DEGs between tumour 1 and tumour 2. (**j**) GSEA showed related pathways enriched in different tumour subtypes. (**k**) Potential gene regulatory network of estrogen response pathway (early and late). (**l**) Expression levels of genes shared by estrogen response pathways (SEMA3B and TFF1) in different tumour subtypes. (**m**) Survival curves of SEMA3B and TFF1 genes in TCGA breast cancer database.

to evaluate the similarity of overall functions between matched domains, we utilized four quadrant diagram with normalized enrichment score (NES) values of the HALLMARK gene sets. The first and third quadrant indicate similar regulation pattern between two domains. Taken domain 5 of P55 and corresponding domain 4 of P59 as an example, HALLMARK processes reaching significance shared similar pattern in both domains (Fig. 6f). Correlation analysis further echoed this result. Domain 5 of P55 and corresponding domain 4 of P59 exhibited highest association compared to other pairs (Fig. 6g-h). In order to understand the clinical indications, the H&E images of selected domains showed same histological features of infiltrating immune cells in the stroma (Fig. 6i). Functional analysis deciphered immune-activities with enriched pathways associated to 'B cell regulation of

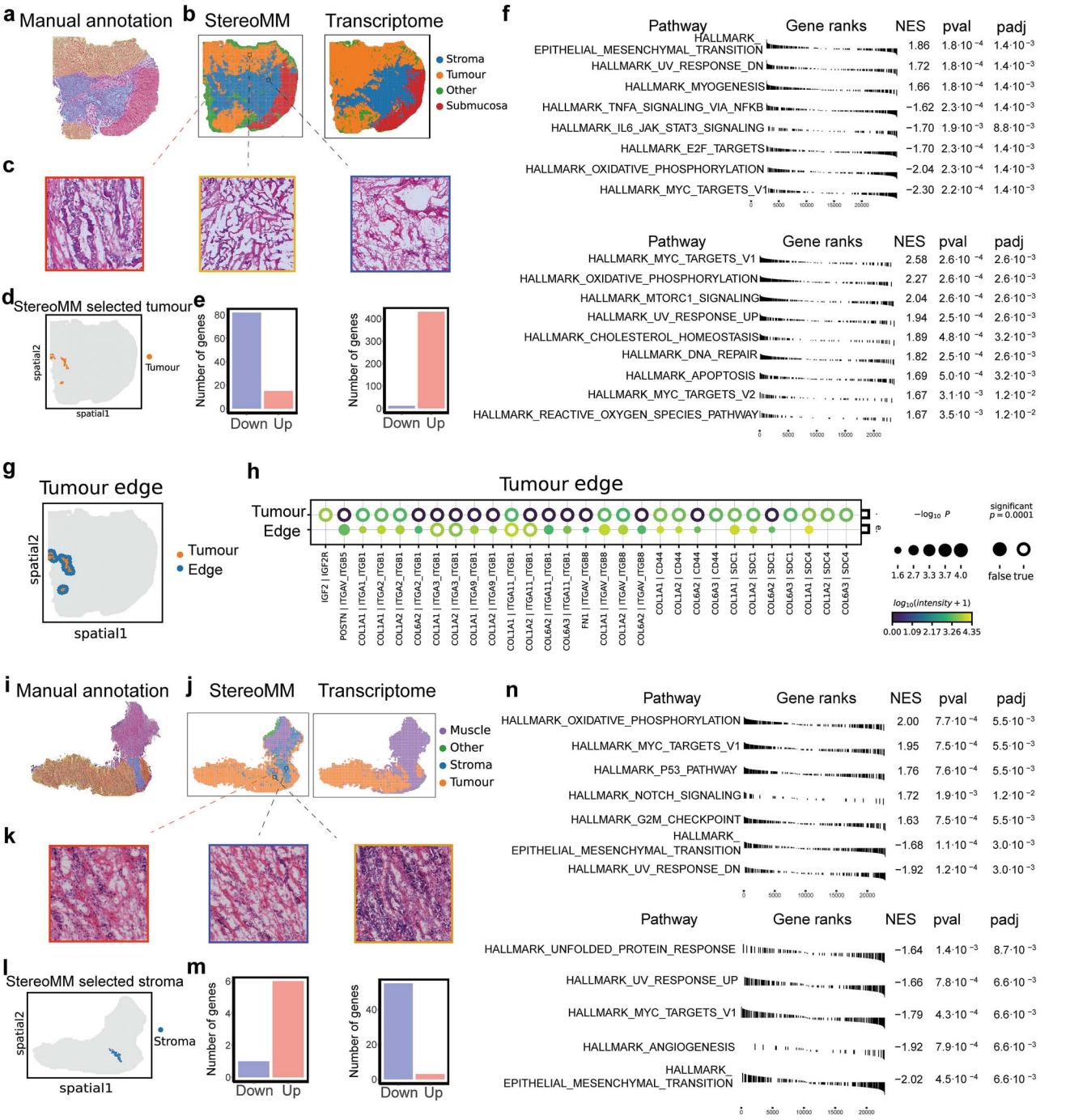


Figure 5. **Overlooked consistent regions detected by StereoMM beyond pathological annotations.** (**a**) Manual pathological annotation based on H&E staining of human colorectal cancer P19 data. Yellow: Lesion, blue: Stroma, red: Mucosa. (**b**) Automatic annotation of spatial domains identified by StereoMM (left) and single transcriptome modality (right). (**c**) Examples of H&E images of different regions of colorectal cancer P19 data. Red: Distinct tumour, yellow: Other tumour, blue: Stroma. (**d**) Spatial location of these distinct tumour regions. (**e**) Bar plot of differential gene counts. Left: Distinct tumour region compared with other tumour regions. Right: Distinct tumour regions compared with stroma regions. (**f**) GSEA showed enrichment in related pathways. Top: Distinct tumour regions compared with other tumour regions. Bottom: Distinct tumour regions compared with stroma regions. (**g**) Spatial location of distinct tumour regions and surrounding stroma. (**h**) Ligand-receptor pairs mediating interactions between distinct tumour and surrounding area. (**i**) Manual pathological annotation based on H&E staining of human colorectal cancer P36 data. Yellow: Lesion, blue: Stroma, purple: Muscle. (**j**) Automatic annotation of spatial domains identified by StereoMM (left) and single transcriptome modality (right). (**k**) Examples of H&E images of different regions of colorectal cancer P19 data. Red: Distinct stroma, yellow: Tumour, blue: Stroma. (**l**) Spatial location of distinct stroma region. (**m**) Bar plot of differential gene counts. Left: Distinct stroma region compared with other stroma regions. Right: Distinct stroma region compared with tumour regions. (**n**) GSEA showed related pathways enriched. Top: Distinct stroma region compared with other stroma regions. Bottom: Distinct stroma region compared with tumour regions.

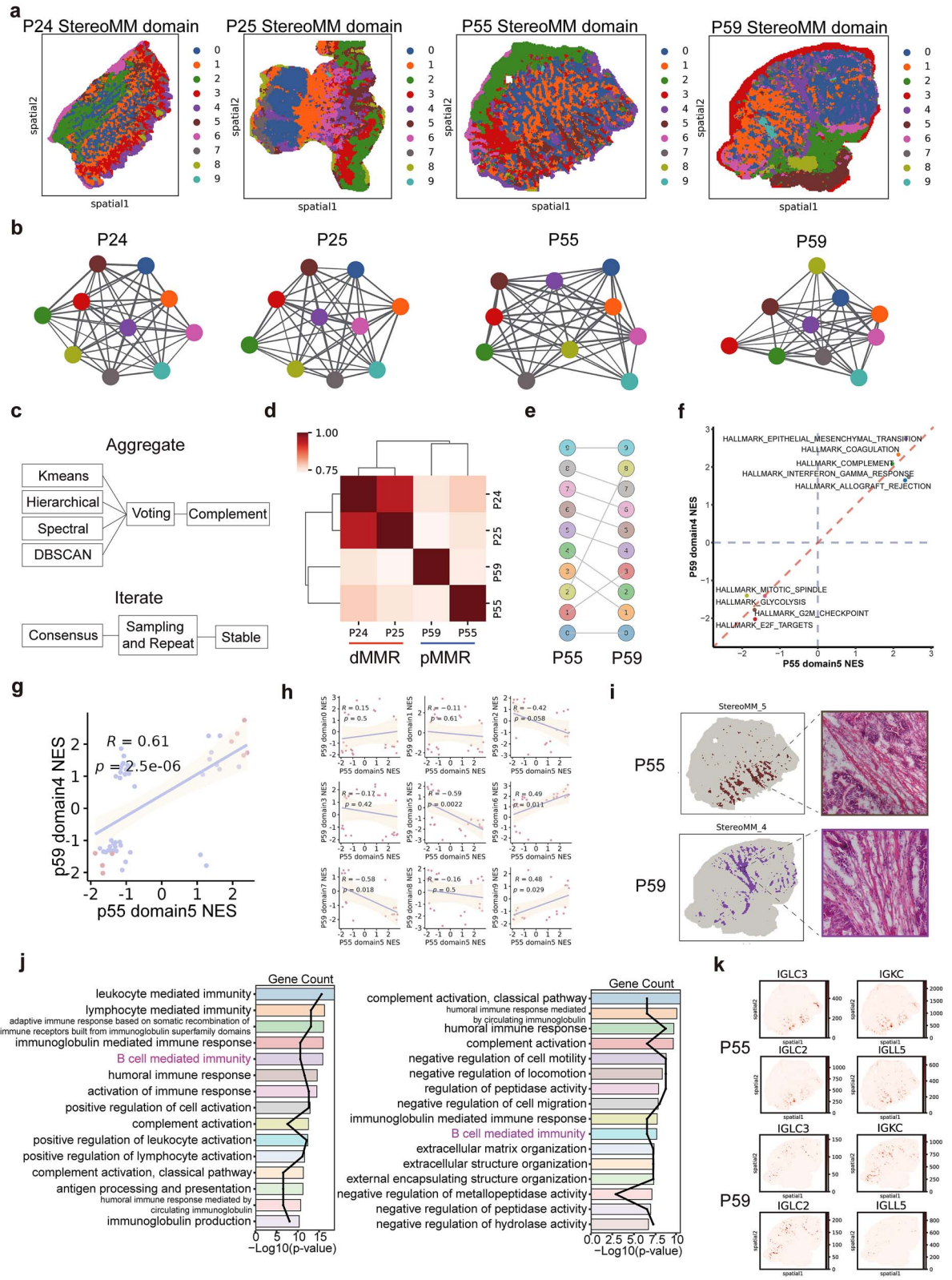


Figure 6. **Graph construction of StereoMM result and patient classification.** (**a**) Spatial domains of CRC cohort identified by StereoMM. (**b**) Patient-level graph constructed by StereoMM domains. (**c**) Flowchart summarizing multiple clustering methods for patient classification. (**d**) Similarity heatmap and classification results of CRC cohort samples. (**e**) Node matching graph of P55 and P59 samples. (**f**) Scatter plot of NES values of domain5 in P55 and domain 4 in P59, divided into four quadrants according to (0, 0). (**g**) Scatter plot of correlation of NES values between domain 5 in P55 sample and domain 4 in P59 sample, fitting regression line. (**h**) Scatter plot of correlation of NES values between domain 5 in P55 sample and other domains in P59 sample, fitting regression line. (**i**) Examples of H&E images of domain5 in P55 and domain 4 in P59. (**j**) Bar plot of GO enrichment analysis. Left: Domain 5 in P55. Right: Domain 4 in P59. (**k**) Spatial location of the expression of immunoglobulin family genes.

immune response' (Fig. 6j) and highly expressed immunoglobulin (IG) family of genes in matched domains of two patients (Fig. 6k). The analysis has been repeated in matched tumoural and immune domains in patient 55 and 59 (Supplementary Fig. 13-14) to validate the observation.

In conclusion, our approach leverages StereoMM to construct a comprehensive graph representation of patient derived tumour microenvironments, to effectively classify patients into distinct groups. The node matching results of domains were proven to be powerful to identify significantly featured domains. Therefore, StereoMM holds the potential to explore multimodal based domains with clinic-pathogenesis associated functions and pave the ways for constructing new patient stratification system.

## Discussion

As spatial transcriptomic (ST) technologies develop, integration with other data modalities provides opportunities for better tissue characterization [29]. Integration of spatial transcriptomic data with conventional hematoxylin and eosin (H&E) histopathology images of tumour tissues opens new avenues for clinical applications [30–33]. The multi-channel images provided in ST contain rich information, including cell morphology and cell status. Changes in morphology may predict cell fate or state even before they are observed in transcriptome output [34]. Meanwhile, spatial relationships between cells can reveal how different cell types and genetic programs relate to each other and their surroundings [35].

Our study introduces StereoMM, a deep learning approach that integrates multimodal data—including gene expression, high-content H&E images, and spatial information—to comprehensively identify tumour subpopulations, significantly advancing beyond conventional methods by considering both spatial and modality interactions within tissue samples. StereoMM employs a cross-attention mechanism facilitating deeper interactions between the modalities across distances, followed by the aggregation of the multimodal features from adjacent tissues using a graph convolutional network. This methodology affords StereoMM with exceptional adaptability and computational efficiency. The attention mechanism enables the model to understand and connect information across different modalities, even when they are far apart. This capability of balancing modality weight bridges the gap between complex algorithmic decisions and human understanding, enabling people to comprehend the principles behind specific predictions or classifications made by the model. The utility of the attention module in mediating information exchange has been substantiated through ablation experiments. By adjusting various parameters, we have demonstrated the robustness of our model. The model can also be calibrated by the end users based on the samples being studied and the experimental requirements. For instance, tissues with lower inter-regional similarity may benefit from a smaller k-nearest neighbours parameter or fewer graph convolutional layers. Such customization can yield results with greater biological relevance across diverse datasets.

StereoMM can be applied to all spatial transcriptomics technologies, including in situ capture and sequencing-based spatial omics data and has been validated on tumour datasets from Stereo-seq and 10X Visium, exhibiting superior performance in spatial contour identification. Comparative analyses with manual annotations have revealed spatial domains that more accurately reflect the ground truths, and congruence with cell subtype marker genes has indicated subpopulation compositions that correlate with biological functions. The intricate spatial architecture of tumour tissues necessitates a detailed analysis of the tumour microenvironment, which is crucial for comprehending tumour biology, unravelling mechanisms of oncogenesis, and identifying therapeutic targets. The refined subpopulations discerned through StereoMM, in conjunction with multimodal data, appear to capture significant biological variations, including genes implicated in tumour progression and intratumoural heterogeneity.

We developed an innovative approach that utilizes StereoMM for discerning tissue-level heterogeneity and patient stratification. StereoMM integrates multi-dimensional perspective, which is crucial for accurate patient stratification. Our study highlights the efficacy of a graph-based methodology in classifying colorectal cancer patients into distinct dMMR and pMMR cohorts, with the capability to align tumour microenvironments through a node-matching strategy. The aggregation of various clustering methods and the application of consensus clustering demonstrate the stability of our classification results. Such reliability is crucial for clinical applications, where accurate patient categorization can provide valuable insights for treatment decisions and prognostic assessments. Quadrant diagram and correlation analyses of inter-node (domain) pathways have revealed a marked functional consistency across matched domains, confirming the synteny across different domains.

StereoMM has demonstrated its spatial recognition capabilities across three cancer types and provided examples for further analysis of the tumour microenvironment and patient stratification. We are eager to expand these experiments to additional tumour categories. Our current findings show strong concordance with ground truth annotations and revealed regions that were potentially missed by pathologists. However, our objectives were to delve into the enigmas underlying the histological data, searching for molecular aberrations that might account for variations in histopathology and therapeutic outcomes. While StereoMM has been applied to spatial transcriptomic analyses using binning or meshing methods, the rapid evolution of ST technology presents new challenges [36]. The ability to integrate modalities beyond spatial transcriptomics and perform multi-slice analyses broadens its applicability in real-world clinical cohorts. Future investigations will explore these potential enhancements to further refine the functionality of StereoMM.

In summary, StereoMM is an innovative and promising approach utilizing attention mechanisms and graph autoencoders for the analysis of spatial transcriptomic data. It facilitates modality fusion through self-supervised learning in the absence of annotations. StereoMM offers a user-friendly tool for the fusion of multimodal information. Medical researchers can use StereoMM for a number of clinically relevant applications, thereby fostering advancements in patient care and therapeutic approaches. Poised to capitalize on forthcoming advancements in measurement technologies, StereoMM holds the potential to significantly improve precision oncology practices in the context of therapeutic decision-making.

## Materials and methods

### Data description and preprocessing

We used five datasets to test the capability of StereoMM. These datasets included a lung cancer dataset from the Stereo-seq platform, two colorectal cancer datasets also from the Stereo-seq platform, a breast cancer dataset and mouse brain dataset from the 10X platform. The H&E images from all datasets were registered with the transcriptomic coordinates.

We divided the lung cancer dataset it into four slices. Each slice was aggregated in a non-overlapping manner with a resolution of 100*100 DNB (bin100) to generate the initial expression matrix. The stereopy package was used to construct anndata for downstream analysis, and cells with a total count less than 300 were filtered out. The raw data from the 10X platform datasets were downloaded from the database and directly read into anndata using the read_visium function in scanpy [37]. Spots with in_tissue = 0 were filtered out. After obtaining the anndata for all the datasets, the same preprocessing steps were applied to the expression matrices: data normalization steps were performed using scanpy with normalize_total, log1p and scale functions. Unsupervised clustering was performed using the tl module function within the scanpy API.

## The StereoMM model

StereoMM is a self-supervised deep learning model that requires three profiles: gene expression matrices, morphological images and spatial location information. The ability of StereoMM to combine multimodal data is built upon the following two principles: (1) fully integrating features between modalities, (2) faithful preservation of effective information from original modalities when reducing dimensions. The two principles are implemented by the attention module and the variational graph autoencoder (VGAE) module, respectively. The specific training process for StereoMM is as follows:

(1) Generated tiled images based on the center points of each bin or spot. The high-level visual features of each patch ($M = \{M_1, M_2, \cdots, M_n\}$) were learned by a unimodal image feature extractor, with options including CNNs and the feature extraction module of the Clinical Histopathology Imaging Evaluation Foundation [38] pathology foundation model. Each patch was given an $s$-dimensional image feature vector (for ResNet50, $s = 2048$, $x_{mi} = \left[vec\left(M_{i,1}\right), vec\left(M_{i,2}\right), \ldots, vec\left(M_{i,s}\right)\right]$). For the entire slide data, we obtained a morphological feature matrix ($X_m \in R^{n \times s}$). Where $n$ is the number of bins or spots.

(2) For the gene expression matrix, we obtained the features of each bin or spot ($T = \{T_1, T_2, \cdots, T_n\}$) through a unimodal feature extractor (principal component analysis, and highly variable genes or genes with highest variance). The dimensionality of transcriptomic features was kept consistent with the dimensionality of morphological features ($x_{ti} = \left[vec\left(T_{i,1}\right), vec\left(T_{i,2}\right), \cdots, vec\left(T_{i,s}\right)\right]$). For the entire slide data, we obtained a transcriptomic feature matrix ($X_t \in R^{n \times s}$).

(3) Once the transcriptomic features and morphological features were obtained, a cross-attention module was used to fuse the features from the two modalities, ensuring that the dimensions of the output features remain consistent with the initial dimensions. The output feature dimensions obtained from the attention module ($s = 2048$) remain the same as before, but the feature vectors were updated.

(4) Updated transcriptomic and morphological features were concatenated as node representations, which together with the adjacency matrix of SNG constituted the final graphical representation: $G = (V, E)$. Where $v_i \in V$ represents the $i$ th bin or spot, and $i = 1, 2, \cdots, n$ represents all $N$ bins or spots. $e_{ij} \in E$ represents the connection between $v_i$ and $v_j$. The graph was reconstructed using the variational graph autoencoder, and the intermediate latent space was extracted as a new fusion feature representation ($Z \in R^{n \times d}$), where $d$ represents the feature dimension after dimensionality reduction. The new fused feature representation $Z$ was used for downstream analysis tasks focused on clustering.

## Construction of SNG

To aggregate the information of neighbouring points for a specified point, StereoMM generates an undirected neighbour graph called SNG based on spatial coordinates.

To construct the SNG, two optional methods are available: (i) the number of the nearest neighbours and (ii) the radius threshold.

Option 1: number of nearest neighbours.

In this approach, the bins or spots (hereinafter referred to as points) are sorted based on the Euclidean distance between a specified point and all other points. The k nearest points to the specified point are then selected as its neighbours. This method ensures a consistent number of connections for each point, regardless of spatial distribution, making it useful when point density varies or when the exact spatial scale of interactions is less critical.

Option 2: radius threshold.

In this method, all points within a defined Euclidean distance from the specified point are selected as neighbours. It captures interactions within a fixed spatial scale, making it ideal for scenarios where relationships are confined to a specific range, such as cell–cell interactions typically occurring within 200 micrometers.

For all data in this article, a radius threshold was used to construct the SNG. Specifically, for Stereo-seq data with bin100, a radius threshold of 100 was chosen. For 10X Visium data, the distance between each spot was calculated based on the resolution of the full image, and this distance was used as the threshold.

Through the above calculation, assuming $A$ is the adjacency matrix of SNG, then $A_{ij} = A_{ji}$ were assigned a value of 1 when there was a connection between vertex $i$ and vertex $j$, otherwise a value of 0 was assigned.

By using this approach, we can construct a neighbour graph that includes the nearest neighbouring points of a specified point, allowing for the aggregation and analysis of the neighbouring points information. This method helps us better understand and interpret the neighbour relationships and interactions in spatial transcriptomics data.

## Attention module

The attention module is the main component of the StereoMM model. The attention mechanism has the capability to automatically learn the information in the query and selectively focus on important information, thereby enhancing the performance of information interaction. The attention module utilizes three main components: queries (Q), keys (K), and values (V). Here, we implemented a cross-attention mechanism, where the morphology and transcriptome modalities were respectively used as the query to guide the update of the other modality. Notably, our attention mechanism generates an $n \times n$ attention weight matrix for all $n$ points across the entire tissue slice. This design ensures that for any point, even those at long distances, the model can capture inter-modal weights.

For morphology-guide transcriptome attention mechanism, we performed the following operations:

StereoMM utilizes three fully connected layers to map transcriptomic features to key tensors and value tensors, as well as map morphological features to a query matrix.

$$Q_m = X_m W_{Qm}$$

$$K_t = X_t W_{Kt}$$

$$V_t = X_t W_{Vt}$$

where $W_{Qm}, W_{Kt}, W_{Vt} \in R^{n\times d_k}$, and $d_k$ denotes the inner dimension that is specific to each attention layer.

We computed morphology-guide transcriptome attention weights using the following steps:

Step 1: Calculated the inner product to determine the correlation between $Q_m$ from morphology and $K_t$ from transcriptome.

$$simi\left(K_i, Q_m\right) = Q_m \bullet K_t^T$$

Step 2: The similarity was normalized by softmax to obtain the correlation weights between morphology and transcriptome.

$$\alpha_{\mathbf{K}_t,Q_m} = softmax\left(simi\left(K_t, Q_m\right)\right) = \frac{\exp\left(simi\left(K_t, Q_m\right)\right)}{\sum_{i=1}^{n} simi\left(K_t, Q_m\right)}$$

Step 3: Calculated the final attention weight through value and similarity score.

$$Attention\left(Q_m, K_t, V_t\right) = \alpha_{K_t,Q_m} \bullet V_t$$

Finally, we defined the attention layer as:

$$Y = softmax\left(X_m W_{Qm} \bullet \left(X_t W_{Kt}\right)^T\right) \bullet X_t W_{Vt}$$

The process of setting the transcriptome as the query involves similar operations. The key difference is that the query is derived from the transcriptomic feature matrix, while the key and value are derived from the morphological feature matrix.

## Graph variational auto-encoder module

VGAE module was used to obtain an embedded representation of multimodal data that incorporates SNG. The autoencoder structure can be trained by self-supervised methods. A standard autoencoder consists of an encoder and a decoder. Replacing the linear layers of the standard autoencoder with GNN can incorporate spatial information. And variational modifications to standard encoder-decoder architectures can be transformed into probabilistic encoders, improving the performance of generated embeddings.

Given the cell adjacency matrix A and the node feature matrix $X$ ($X \in R^{n\times 2s}$) obtained by concatenating morphological features and transcriptomic features. VGAE learns the latent representation $Z$ ($Z \in R^{n\times d}$) of the cell graph. Then, $Z$ is used to reconstruct the node feature matrix or adjacency matrix in the encoder.

The generative model (encoder) is divided into the following two parts:

Step1: The encoder part of VGAE is parameterized by GNN. The mean ($\mu$) and the logarithm of the variance ($log\sigma$) are obtained from two GNNs. StereoMM provides three options for the GNN architecture, including GCN, GAT, and GraphSAGE. By default, StereoMM uses GCN.

$$\mu = GCN_\mu\left(X, A\right) = \tilde{A}ReLU\left(\tilde{A}XW_{\mu 0}\right)W_{\mu 1}$$

$$log\sigma = GCN_{log\sigma}\left(X, A\right) = \tilde{A}ReLU\left(\tilde{A}XW_{log\sigma 0}\right)W_{log\sigma 1}$$

$$\tilde{A} = D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$$

Where, $W_{\mu i}$ and $W_{log\sigma i}$ represent the weight matrix of $i$ th layer used to infer $\mu$, and $log\sigma$, and $\tilde{A}$ represents the symmetrically normalized adjacency matrix, respectively. By default, the first layer and the second layer are set to 2048 and 512, respectively.

Step2: According to the $\mu$ and the $log\sigma$, obtain the hidden representation $Z$.

$$q\left(Z|X, A\right) = \prod_{i=1}^{N} q\left(z_i|X, A\right)$$

$$q\left(z_i|X, A\right) = N\left(z_i|\mu_i, \mathbf{diag}\left(\sigma_i^2\right)\right)$$

The variational approximation (decoder) reconstructs node features using hidden representation of the same GNN type as the encoder. Taking GCN as an example, the decoder includes $i+1$ layers.

$$\hat{X} = \tilde{A}\left(ReLU\left(\tilde{A}\left(ReLU\left(\tilde{A}ZW_{d0}\right)W_{d1}\right)\right)W_{d2}\right)$$

Where $W_{di}$ represent the weight matrix of $i$ th layer in decoder.

For different task requirements, although it is not the default setting, we also provide a decoder that reconstructs the graph structure.

$$p\left(A|Z\right) = \prod_{i=1}^{N}\prod_{j=1}^{N} p\left(A_{ij}|z_i, z_j\right)$$

$$p\left(A_{ij} = 1|z_i, z_j\right) = \sigma\left(z_i^T z_j\right)$$

where $A_{ij}$ are the elements of A and $\sigma(\cdot)$ is the logistic sigmoid function.

## Loss function

StereoMM uses VGAE to reconstruct graphs and perform self-supervised training. Generally, two loss functions are set during model training to guide the model optimization direction: (i) Reconstruction loss: the discrepancy between the original node features and their reconstructed counterparts, serving as a metric for evaluating the effectiveness of the VGAE in accurately reconstructing node attributes. (ii) Variational Kullback–Liebler (kl) divergence loss: the error between the variational probability distribution and the prior distribution. VGAE uses variational inference to learn the latent variable distribution of the nodes and maximize the lower bound estimation of the node representations. The loss is typically computed using the kl divergence between the learned distribution and the prior distribution of the latent variables. Ultimately forces two distributions to be similar and provides regularity.

(1) Reconstruction loss:

$$L_{recon} = MSE\left(X, \hat{X}\right) = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \hat{X}_i\right)^2$$

Where $\hat{X}$ represents the node feature matrix reconstructed by the decoder of VGAE.

(2) Variational kl divergence loss:

$$L_{vkl} = KL\left(q\left(Z|X, A\right)\right)\left\|p(Z) = \frac{1}{2}\sum_{i=1}^{n}\left(1 + iog\sigma_i^2 - \mu_i^2 - \sigma_i^2\right)\right.$$

The total loss function of the StereoMM model is summarized as:

$$L_{total} = L_{recon} + \beta L_{vkl}$$

Where $\beta$ is a parameter that controls the weight of the two losses, the default is $\frac{1}{n}$. And $n$ is the number of bins or spots.

## StereoMM parameter settings

For the H&E images, we performed segmentation of the entire tissue image after registering it with the transcriptome. Tiling was performed on the registered images based on the centroid coordinates of each point in the transcriptome. This approach ensured that the number of generated image patches matched the number of transcriptomic points. Specifically, for the bin100 data from Stereo-seq, we used patches of size 128*128 pixels (64*64 $\mu$m) corresponding to the centre of the corresponding transcriptomes. Patches of size 200*200 pixels were extracted from the spot data obtained from the 10X datasets, in a similar manner. Morphological features were extracted from each patch using the default parameters of a pre-trained ResNet50 model.

When performing the fusion strategy with StereoMM, consistent parameters were used for all the datasets: GNN with GCN architecture and a hidden layer structure of 2048_512_100.

## Spatial domain detection

After applying StereoMM to analyse ST data, we learned accurate low-dimensional latent representations (Z) that represent the multimodal fused features of each bin or spot. StereoMM uses Leiden algorithm to cluster the latent representations of the bins or spots. In the Leiden algorithm, the graph is typically constructed by connecting each node to its k nearest neighbours. The selection of neighbours is crucial as it defines the local neighbourhood structure used for clustering. For fair comparison, we took the same default 20 nearest neighbours for all data to build the graph. We adjusted the value of resolution to determine the final number of clusters. These clustered groups were defined as spatial domains.

## Spatial domain evaluation

We employed multiple different metrics to evaluate the clustering results of spatial domains. These metrics include both internal and external indicators of clustering quality, as well as measures that assess spatial diversity.

## Clustering external indicators

For data where ground truth labels were available, we used external clustering metrics for evaluation. ARI and NMI are widely used external clustering metrics to evaluate the consistency between clustering results and external labels or reference information.

$$AdjustedIndex = \frac{Index - ExpectedIndex}{MaxIndex - ExpectedIndex}$$

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}$$

ARI measures the similarity between clustering results and external labels. It takes into account the impact of randomness on the Rand Index and adjusts it to have a range between −1 and 1. A higher ARI value indicates a higher consistency between the clustering results and external labels.

$$NMI(X;Y) = \frac{I(X,Y)}{F(H(X), H(Y))}$$

$$I(X;Y) = \sum_x \sum_y p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

NMI is a standardized variant of Mutual Information that scales the output to a range between 0 to 1. Higher ARI and NMI values indicate a higher consistency between the clustering results and external labels. We calculated ARI and NMI using functions from the python package, sklearn.metrics.

## Clustering internal indicators

For data without ground truth, the internal clustering metrics such as Calinski-Harabasz Index (CH), Davies-Bouldin Index (DB), and Silhouette Coefficient (SC) were used to evaluate the quality of clustering results.

CH index measures the compactness within clusters and separation between clusters. Higher CH values indicate better clustering results with tighter and more separated clusters. The silhouette coefficient is similar to CH. A higher value indicates better clustering results, with good separation and strong internal cohesion. While DB Index is the opposite. The DB Index is computed by averaging the ratios of within-cluster scatter to between-cluster separation for each cluster, with lower DB values indicating better clustering results. We calculated CH, DB and silhouette coefficient using functions from the python package, sklearn.metrics.

## Spatial distribution indicator

Local inverse Simpson's index (LISI) is a biodiversity metric that measures diversity and evenness within a specific area. It is the reciprocal of the Simpson's index. A higher value of the index indicates a more diverse and evenly distributed community, suggesting a higher level of spatial continuity. This index helps assess the spatial patterns of species distribution and identify areas important for maintaining ecological connectivity. We calculated LISI by writing a function using a python script available at https://github.com/STOmics/StereoMMv1.

**Key Points**

- StereoMM is a cutting-edge deep learning tool designed to integrate diverse omics data—including gene expression, histological images, and spatial information—using a graph-based fusion model enhanced with a cross-attention mechanism.
- StereoMM demonstrates the potential for scientific discovery by achieving superior accuracy in identifying spatial domains compared to competing software, excelling in distinguishing heterogeneous tumour regions and interrogating spatial molecular mechanisms associated with prognosis to provide novel insights into complex biological systems.
- Beyond basic research, StereoMM offers significant clinical value by refining spatial multimodal features through patient graphing to enable precise classification of colorectal cancer patients by mismatch repair status, addressing other critical clinical challenges, and showcasing its transformative potential across both scientific and clinical applications.

## Acknowledgements

The author acknowledges Mr. Zidong Su (BGI Research Institute, Shenzhen) for his assistance in the design of the theoretical framework.

## Author contributions

Conceptualization: Bingying Luo, Jiajun Zhang, Weimiao Yu, Xun Xu, Ao Chen, and Fei Teng.

Supervision: Jiajun Zhang, Weimiao Yu, Jinmiao Chen, Xun Xu, Ao Chen, Sha Liao, Xi Feng, and GuiBo Li.

Software: Bingying Luo and Fei Teng.

Data curation, Formal analysis, Resources: Bingying Luo, Fei Teng, Jiajun Zhang, WeiXuan Chen, Mei Li, Xuanzhu Liu, Huaqiang Huang, Yu Feng, Xing Liu, Min Jian, Xue Zhang.

Methodology: Bingying Luo, Xuanzhu Liu.

Writing—original draft: Bingying Luo, Weimiao Yu, Guo Tang, Jiajun Zhang, Fei Teng.

Visualization: Bingying Luo, Fei Teng, and Weimiao Yu.

Writing—review & editing: Jiajun Zhang, XunXu, Weimiao Yu, Jinmiao Chen, Feng Xi, Xing Liu, Guibo Li, Qu Chi, Xin Liu.

Project administration: Jiajun Zhang, Weimiao Yu, Fei Teng, Jinmiao Chen, Sha Liao, and Ao Chen.

Investigation: Bingying Luo, Jinmiao Chen, Guo Tang, Jiajun Zhang, and Fei Teng.

## Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: The authors declare no competing interests.

## Funding

This study is supported by the Science and Technology Innovation Key R&D Program of Chongqing (CSTB2023TIAD-STX0002).

## Data availability

Lung adenosquamous carcinoma Stereo-seq data was downloaded from GSA-human (https://ngdc.cncb.ac.cn/gsa-human/) under Project HRA004240. Colorectal cancer Stereo-seq data were downloaded from CNGB Nucleotide Sequence Archive (CNSA: http://db.cngb.org accession number CNP0002432) and China National GenBank Database (accession number: STT0000036, https://db.cngb.org/stomics/project/STT0000036). The DLPFC dataset [39] is downloaded from the spatialLIBD package (http://spatial.libd.org/spatialLIBD). 10X Visium mouse brain data and breast cancer data were downloaded from 10x Genomic-Datasets (https://www.10xgenomics.com/datasets), the specific dataset names are V1_Mouse_Brain_Sagittal_Posterior (MouseBrainSerialSection2(Sagittal-Posterior)-10xGenomics) and V1_Breast_Cancer_Block_A_Section_1 (Human Breast Cancer: Ductal Carcinoma In Situ, Invasive Carcinoma (FFPE) - 10x Genomics) respectively.

## Code availability

Code for data analysis is available at https://github.com/STOmics/StereoMMv1.

## Declaration of interests

Some parts of this study are covered by a patent application, with the application number [PCT/CN2023/140756]. The patent is held by [BGI Research, Chongqing]. The authors declare that there are no other conflicts of interest that could influence the results of this study.

## Ethics approval and consent of participate

This study does not require ethics approval or informed consent from participants.

## Consent for publication

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

## References

1. Johnson DB, Bordeaux J, Kim JY. *et al.* Quantitative spatial profiling of PD-1/PD-L1 interaction and HLA-DR/IDO-1 predicts improved outcomes of anti-PD-1 therapies in metastatic melanoma" (in eng). *Clin Cancer Res* 2018;**24**:5250–60. https://doi.org/10.1158/1078-0432.Ccr-18-0309
2. Wang Y, Song B, Wang S. *et al.* Sprod for de-noising spatially resolved transcriptomics data based on position and image information. *Nat Methods* 2022;**19**:950–8. https://doi.org/10.1038/s41592-022-01560-w
3. Williams CG, Lee HJ, Asatsuma T. *et al.* An introduction to spatial transcriptomics for biomedical research. *Genome Med* 2022;**14**:68. https://doi.org/10.1186/s13073-022-01075-1
4. Zhu X, Li X, Ong K. *et al.* Hybrid AI-assistive diagnostic model permits rapid TBS classification of cervical liquid-based thin-layer cell smears. *Nat Commun* 2021;**12**:3541. https://doi.org/10.1038/s41467-021-23913-3
5. Zhou Y, Chia MA, Wagner SK. *et al.* A foundation model for generalizable disease detection from retinal images. *Nature* 2023;**622**:156–63. https://doi.org/10.1038/s41586-023-06555-x
6. Bulten W, Kartasalo K, Chen PHC. *et al.* Artificial intelligence for diagnosis and Gleason grading of prostate cancer: the PANDA challenge. *Nat Med* 2022;**28**:154–63. https://doi.org/10.1038/s41591-021-01620-2
7. Huo X, Ong KH, Lau KW. *et al.* A comprehensive AI model development framework for consistent Gleason grading. *Commun Med* 2024;**4**:84. https://doi.org/10.1038/s43856-024-00502-1
8. Pham D, Tan X, Balderson B. *et al.* Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nat Commun* 2023;**14**:7739. https://doi.org/10.1038/s41467-023-43120-6
9. Hu J, Li X, Coleman K. *et al.* SpaGCN: integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network, (in eng). *Nat Methods* 2021;**18**:1342–51. https://doi.org/10.1038/s41592-021-01255-8
10. Dong K, Zhang S. deciphering spatial domains from spatially resolved transcriptomics with an adaptive graph attention auto-encoder, (in eng). *Nat Commun* 2022;**13**:1739. https://doi.org/10.1038/s41467-022-29439-6
11. Bao F, Deng Y, Wan S. *et al.* Integrative spatial analysis of cell morphologies and transcriptional states with MUSE. *Nat Biotechnol* 2022;**40**:1200–9. https://doi.org/10.1038/s41587-022-01251-z
12. Xu H, Fu H, Long Y. *et al.* Unsupervised spatially embedded deep representation of spatial transcriptomics. *Genome Med* 2024;**16**:12. https://doi.org/10.1186/s13073-024-01283-x

13. Wang L, Hu Y, Xiao K. *et al.* Multi-modal domain adaptation for revealing spatial functional landscape from spatially resolved transcriptomics. *Brief Bioinform* 2024;**25**:bbae257. https://doi.org/10.1093/bib/bbae257
14. Chen A, Liao S, Cheng M. *et al.* spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays, (in eng). *Cell* 2022;**185**:1777–1792.e21. https://doi.org/10.1016/j.cell.2022.04.003
15. Lein ES, Hawrylycz MJ, Ao N. *et al.* genome-wide atlas of gene expression in the adult mouse brain, (in eng). *Nature* 2007;**445**:168–76. https://doi.org/10.1038/nature05453
16. Zhao R, Xu Y, Chen Y. *et al.* clonal dynamics and stereo-seq resolve origin and phenotypic plasticity of adenosquamous carcinoma, (in eng). *NPJ Precis Oncol* 2023;**7**:80. https://doi.org/10.1038/s41698-023-00430-8
17. Li Q, Wang R, Yang Z. *et al.* Molecular profiling of human non-small cell lung cancer by single-cell RNA-seq. *Genome Med* 2022;**14**:87. https://doi.org/10.1186/s13073-022-01089-9
18. Roy S, Sunkara RR, Parmar MY. *et al.* EMT imparts cancer stemness and plasticity: new perspectives and therapeutic potential, (in eng). *Front Biosci (Landmark Ed)* 2021;**26**:238–65. https://doi.org/10.2741/4893
19. Grasset EM, Dunworth M, Sharma G. *et al.* Triple-negative breast cancer metastasis involves complex epithelial-mesenchymal transition dynamics and requires vimentin. *(in eng), Sci Transl Med* 2022;**14**:eabn7571. https://doi.org/10.1126/scitranslmed.abn7571
20. Cui J, Zhang C, Lee JE. *et al.* MLL3 loss drives metastasis by promoting a hybrid epithelial-mesenchymal transition state, (in eng). *Nat Cell Biol* 2023;**25**:145–58. https://doi.org/10.1038/s41556-022-01045-0
21. Oshi M, Takahashi H, Tokumaru Y. *et al.* The E2F pathway score as a predictive biomarker of response to Neoadjuvant therapy in ER+/HER2- breast cancer(in eng). *Cells* 2020;**9**:1643. https://doi.org/10.3390/cells9071643
22. Schuhwerk H, Brabletz T. mutual regulation of TGF$\beta$-induced oncogenic EMT, cell cycle progression and the DDR, (in eng). *Semin Cancer Biol* 2023;**97**:86–103. https://doi.org/10.1016/j.semcancer.2023.11.009
23. Castro-Rivera E, Ran S, Thorpe P. *et al.* Semaphorin 3B (SEMA3B) induces apoptosis in lung and breast cancer, whereas VEGF165 antagonizes this effect, (in eng). *Proc Natl Acad Sci U S A* 2004;**101**:11432–7. https://doi.org/10.1073/pnas.0403969101
24. Yi J, Ren L, Li D. *et al.* Trefoil factor 1 (TFF1) is a potential prognostic biomarker with functional significance in breast cancers. *(in eng), Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* 2020;**124**:109827. https://doi.org/10.1016/j.biopha.2020.109827
25. Andersson A, Larsson L, Stenbeck L. *et al.* Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions. *Nat Commun* 2021;**12**:6012. https://doi.org/10.1038/s41467-021-26271-2
26. Zhang R, Feng Y, Ma W. *et al.* Spatial transcriptome unveils a discontinuous inflammatory pattern in proficient mismatch repair colorectal adenocarcinoma. *Fundamental Research* 2023;**3**:640–6. https://doi.org/10.1016/j.fmre.2022.01.036
27. Kasprzak A, Adamek A. insulin-like growth factor 2 (IGF2) Signaling in colorectal cancer-from basic research to potential clinical applications, (in eng). *Int J Mol Sci* 2019;**20**:4915. https://doi.org/10.3390/ijms20194915
28. Zhang Z, Wang Y, Zhang J. *et al.* COL1A1 promotes metastasis in colorectal cancer by regulating the WNT/PCP pathway, (in eng). *Mol Med Rep* 2018;**17**:5037–42. https://doi.org/10.3892/mmr.2018.8533
29. Rao A, Barkley D, França GS. *et al.* exploring tissue architecture using spatial transcriptomics, (in eng). *Nature* 2021;**596**:211–20. https://doi.org/10.1038/s41586-021-03634-9
30. Vandereyken K, Sifrim A, Thienpont B. *et al.* methods and applications for single-cell and spatial multi-omics, (in eng). *Nat Rev Genet* 2023;**24**:494–515. https://doi.org/10.1038/s41576-023-00580-2
31. Wu Y, Cheng Y, Wang X. *et al.* spatial omics: navigating to the golden era of cancer research, (in eng). *Clin Transl Med* 2022;**12**:e696. https://doi.org/10.1002/ctm2.696
32. Tran KA, Kondrashova O, Bradley A. *et al.* deep learning in cancer diagnosis, prognosis and treatment selection, (in eng). *Genome Med* 2021;**13**:152. https://doi.org/10.1186/s13073-021-00968-x
33. Boehm KM, Khosravi P, Vanguri R. *et al.* Harnessing multimodal data integration to advance precision oncology. *Nat Rev Cancer* 2022;**22**:114–26. https://doi.org/10.1038/s41568-021-00408-3
34. Buggenthin F, Buettner F, Hoppe PS. *et al.* Prospective identification of hematopoietic lineage choice by deep learning. *Nat Methods* 2017;**14**:403–6. https://doi.org/10.1038/nmeth.4182
35. Kleino I, Frolovaitė P, Suomi T. *et al.* Computational solutions for spatial transcriptomics. *Comput Struct Biotechnol J* 2022;**20**:4870–84. https://doi.org/10.1016/j.csbj.2022.08.043
36. Bressan D, Battistoni G, Hannon GJ. the dawn of spatial omics, (in eng). *Science* 2023;**381**:eabq4964. https://doi.org/10.1126/science.abq4964
37. Wolf FA, Angerer P, Theis FJ. SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biol* 2018;**19**:15. https://doi.org/10.1186/s13059-017-1382-0
38. Wang X, Zhao J, Marostica E. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* 2024;**634**:970–8. https://doi.org/10.1038/s41586-024-07894-z
39. Maynard KR, Collado-Torres L, Weber LM. *et al.* Transcriptome-scale spatial gene expression in the human dorsolateral prefrontal cortex. *Nat Neurosci* 2021;**24**:425–36. https://doi.org/10.1038/s41593-020-00787-0