SpaFoundation: a visual foundation model for spatial transcriptomics

Ning Zhang[1,#], Yahui Long[1,#], Sibo Xia[2], Pengzhen Jia[1], Yujie Jin[1], Siqi Chen[1], Jinmiao Chen[3], Min Li[1,*]

[1]School of Computer Science and Engineering, Central South University, Changsha 410083, China

[2]School of Automation, Central South University, Changsha 410083, China

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

Singapore 138648, Singapore.

[#]These authors contributed equally.

[*]Corresponding author. Email: limin@mail.csu.edu.cn.

Abstract

Spatial transcriptomics has revolutionized biological research but is limited by trade-offs between spatial resolution and gene coverage. Here we present SpaFoundation, a versatile visual foundational model with 80 million trainable parameters, pre-trained on 1.84 million histological image patches to learn general-purpose imaging representations. Using histology images alone, SpaFoundation accurately infers spatial gene expression and demonstrates strong robustness and transferability across tasks such as super-resolution inference, tumor detection, and spatial clustering.

Main

Spatial transcriptomics (ST) has emerged as a powerful technology for capturing both gene expression and spatial localization information within tissues[1–4], enabling researchers to characterize spatial gene expression patterns, study cell-cell communications, and resolve the spatiotemporal order of cellular development[5–8]. ST technologies can be roughly divided into two categories, including sequencing-based technologies, such as STARmap[9], 10x Visium, SLIDE-seqV2[10], and Stereo-seq[11], and imaging-based technologies, such as DNA seqFISH+[12], MERSCOPE[13], and Nanostring CosMx[14]. These methods differ in their spatial resolution and gene coverage. Sequencing-based methods are capable of measuring genome-wide gene expression but face lower spatial resolution. By contrast, Imaging-based methods have higher spatial resolution but are limited to lower throughput. These limitations constraint their broad applicability. Alternatively, computational approaches can bridge the gap between spatial resolution and gene coverage by inferring genome-wide expressions from histological images[15,16].

Histological images and gene expression profiles are intrinsically linked and both

reflect the underlying biological state and organization of tissues, suggesting the possibility to predict gene expressions from histological images[17,18]. There exist several computational approaches that have demonstrated considerable advances in this field, including His2ST[19], His2Gene[20], THItoGene[21], BLEEP[22], and TRIPLEX[23]. Typically, these existing methods use convolutional neural network (CNN)[24] and Transformer[25] architectures to extract vision features around captured spots from histological image patches. These learned features are then utilized to capture biologically relevant gene patterns. Despite promising results, conventional imaging-based prediction methods still suffer from limitations when transferred to real-world applications. First, existing methods are typically trained on specific tasks and datasets, resulting in poor generalization across different tasks, organs, or platforms. Furthermore, these methods often struggle to capture complex relationships between tissue morphology and gene expression patterns due to constrained model size and limited training data. Foundation models have recently achieved successful applications in computational pathology. Representative tools such as UNI[26], CONCH[27], CHIEF[28], use large-scale whole-slide images or image-caption pairs to pretrain visual or visual-language foundation models that demonstrate robust performance in downstream vision and visual-language tasks. However, these methods cannot be directly applicable to spatial transcriptomics due to inherent differences in image characteristics.

In this study, we present SpaFoundation (Fig. 1a), a visual foundation model to extract general-purpose histological imaging features for spatial transcriptomics analysis. Building on a teacher–student architecture, SpaFoundation incorporated self-distillation and masked image modeling (MIM) to enhance the learning of high-level semantic and local structural features. We pre-trained SpaFoundation on HEST-1K-3 dataset (Fig. 1b), which comprises 1.84 million image patches from 1,113 samples across 26 tissue organs. SpaFoundation outperformed traditional methods with minimal or zero-shot fine-tuning across four downstream tasks, including spatial gene expression prediction, high-resolution gene expression inference, tumor detection, and spatial clustering (Fig. 1c). In addition, we conducted an ablation study to demonstrate the effectiveness of the foundation model in extracting histological imaging features (full experimental details are in the Supplementary Information).

SpaFoundation accurately predicts spatial gene expression profiles with histological images alone

To evaluate the performance of SpaFoundation, we first applied the model to predict spatial gene expression from histological images and compared it against five state-of-the-art methods, including HistoGene[20], His2ST[19], THItoGene[21], BLEEP[22], and TRIPLEX[23]. We used the cSCC dataset[29], a human skin dataset comprising 12 squamous cell carcinoma samples from 4 patients. The model was tested for both highly expressed gene (HEG) prediction and highly variable gene (HVG) prediction.

For the HEG prediction, we followed the protocol used in the TRIPLEX[23] and selected the top 250 highly expressed genes as prediction targets. Visually, SpaFoundation's predictions more closely match the ground truth compared to HistoGene, His2ST, THItoGene, BLEEP, and TRIPLEX (Fig. 1d and Supplementary

Fig. S1a). To quantitatively assess effectiveness, we calculated the Pearson correlation coefficient (PCC) to measure the consistency between the predictions and the ground truth. SpaFoundation consistently outperformed baseline methods (Fig. 1e and Supplementary Fig. S1b), showcasing high agreement with visual results. To further validate robustness, we examined results across all 12 samples, where SpaFoundation performed better again than competing methods (Fig. 1f and Supplementary Fig. S1c). For the HVG prediction, we followed the His2ST[19] protocol and selected the top 1,000 highly variable genes. SpaFoundation achieved higher PCC scores than competing methods in most cases (Fig. 1g and Supplementary Fig. S2).

Notably, baseline methods exhibited varying sensitivity to the two prediction tasks. For example, His2ST performed better on HVG prediction than on HEG, whereas BLEEP exhibited better on HEG than HVG. In contrast, SpaFoundation demonstrated stable and strong performance across both tasks, underscoring its robustness. We further validated SpaFoundation on an independent HER2+ dataset[30], where it again achieved competitive results (Supplementary Figs. S3-5), supporting the generalizability and utility of the model.

SpaFoundation enables accurate inference of high-resolution gene expression from low-resolution histological images

Next, we assessed SpaFoundation's effectiveness in inferring high-resolution spatial gene expression from low-resolution spatial transcriptomics data. Spatial transcriptomics (ST) technologies often require trade-offs between spatial resolution and transcriptome coverage. Imaging-based ST platforms offer high spatial resolution but typically capture only a limited number of genes, while sequencing-based platforms cover more genes but at lower spatial resolution. To address this limitation, a promising alternative is to infer gene expression in silico using histological images.

For evaluation, we utilized a human breast cancer Visium HD dataset that profiles gene expression at two spatial resolutions, i.e., 16μm and 8μm (Fig. 2a). We fine-tuned the pre-trained SpaFoundation model on 16μm-resolution patches (~140,000 spots), and tested its effectiveness on the higher-resolution 8μm data (~530,000 spots). We compared SpaFoundation to ST-Net[17] a state-of-the-art method for high-resolution gene prediction. Visually, SpaFoundation more closely recapitulated the ground truth gene expression patterns than ST-Net (Fig. 2b and Supplementary Fig. S6). Quantitatively, SpaFoundation achieved a PCC score of 0.84, improving ST-Net (0.65) by 29.23% (Fig. 2c). Next, we examined both methods on the top 1,000 HVGs and two marker genes, where SpaFoundation attained a median PCC score of 0.19, representing a 91.28% improvement over ST-Net (Fig. 2d). Focusing on the top 50 best-predicted genes, SpaFoundation reached a median PCC score of 0.60, again outperforming ST-Net (Fig. 2d). These findings underscored SpaFoundation's superior capability in inferring high-resolution spatial gene expression from low-resolution ST data.

SpaFoundation accurately identifies tumor regions without gene expression

data

To further assess SpaFoundation's ability to extract biologically meaningful features from tissue morphology, we applied it to identify tumor detection in spatial tissue images. For this evaluation, we used a dataset from He et al. (2020) containing 68 tissue sections from 23 breast cancer patients. The pre-trained SpaFoundation model was fine-tuned on this dataset for one epoch, with five patients held out for testing and the remaining used for training. We compared SpaFoundation against ResNet50[31] and VGG16[32], two widely used convolutional neural networks that were pretrained in a supervised manner on the large-scale natural image dataset ImageNet-1K. These two methods are commonly employed across various computer vision tasks.

Visually, the tumor segmentation maps generated by SpaFoundation showed greater alignment with ground truth annotations, whereas both ResNet50 and VGG16 frequently misclassified non-tumor regions as tumor regions (Fig. 2e). To quantitatively assess performance, we calculated accuracy (ACC), F1 score (F1), and area under ROC curve (AUC) using manual annotations. SpaFoundation consistently outperformed ResNet50 and VGG16 in terms of both ACC and F1 (Fig. 2f, g). SpaFoundation achieved an AUC of 0.94 across all test samples, outperforming ResNet and VGG by 4.16% and 1.55%, respectively (Fig. 2h). These results highlight the advantages of SpaFoundation over conventional vision networks. On one hand, they demonstrate the strong capability of pretrained foundation models to learn biologically meaningful features from histological images. On the other hand, SpaFoundation, as a domain-specific and unsupervised pretrained model, outperforms models trained in a supervised manner on general-purpose image data, underscoring the necessity of developing large-scale foundation models tailored to domain-specific biomedical applications. Overall, these findings confirm the scalability and effectiveness of SpaFoundation for tumor detection in spatial tissue images.

SpaFoundation enables accurate spatial clustering using only image features

Finally, we further validated SpaFoundation's broad applicability to spatial clustering. For better evaluation, we used eight manually annotated tissue sections from the HER2+ dataset, and compared SpaFoundation with state-of-the-art methods, including Scanpy[33], SpaGCN[34], and stLearn[35]. This task was designed as a zero-shot scenario where spot-level imaging representations were extracted from the pretrained SpaFoundation model without any fine-tuning. These representations were then subjected to PCA for dimensionality reduction, followed by Leiden clustering.

We present the clustering results for samples H1 and C1 (Fig. 2i). For the H1 sample, which was manually annotated into six regions, including invasive cancer, breast glands, in situ cancer, adipose tissue, immune infiltrate, and connective tissue, both SpaFoundation and stLearn more accurately identified invasive and in situ cancer regions with clear boundaries compared to Scanpy and SpaGCN. While Scanpy distinguished cancer from non-cancer regions, it failed to separate invasive from in situ cancer, and SpaGCN also mixed these two cancer types. Among non-cancer regions, SpaFoundation was the only method that accurately delineated adipose tissue with minimal noise. In the C1 sample, annotated into three regions, namely invasive cancer,

adipose tissue, and connective tissue, we constrained all methods to produce the same number of clusters as the ground truth. SpaFoundation yielded spatial segmentations that align most closely with expert-defined histological annotations, while Scanpy, SpaGCN, and stLearn produced mixed or noisy boundaries. Notably, SpaFoundation clearly separated invasive cancer from connective tissue regions, demonstrating its ability to capture biologically relevant structures based solely on tissue morphology.

For quantitative assessment, we utilized two widely used metrics, i.e., Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI). SpaFoundation outperformed all competing methods, achieving ARI and NMI scores of 0.4219 and 0.5309 on sample H1, and 0.6058 and 0.4687 on sample C1, respectively (Fig. 2j,k). To assess robustness, we tested SpaFoundation and competing methods on eight samples and evaluated their performance using ARI and NMI. SpaFoundation achieved the highest median ARI and NMI scores of 0.3176 and 0.3761, improving the second best-performing methods by 29.94% and 30.77%, respectively, thereby confirming its strong generalizability across diverse samples (Fig. 2l). By examining the clustering results on additional samples (Supplementary Figs. S7,8), SpaFoundation exhibits superior clustering performance on the majority of samples, validating its effectiveness for image-based spatial clustering.

Conclusion

In summary, we present SpaFoundation, a vision foundational model with 80 million trainable parameters for the analysis of spatial transcriptomics (ST). SpaFoundation has important implications for real-world applications, particularly in addressing the limitations of current ST technologies, which are often constrained by high cost, limited resolution, and low throughput. Notably, by leveraging histological images alone, SpaFoundation enables the inference of spatial gene expression profiles, identification of tumor regions, and delineation of spatial domains, offering a cost-efficient and scalable alternative to experimental profiling. The key advantage of SpaFoundation lies in its ability to learn general-purpose, morphology-aware representations through large-scale pretraining on 1.84 million histological images. This enables strong performance across diverse downstream ST analysis tasks without the need for extensive fine-tuning. Through analysis of various downstream applications, SpaFoundation demonstrates remarkable transferability and robustness across multiple tissues and scenarios. Moreover, we have shown that SpaFoundation consistently outperforms conventional methods trained on limited, task-specific data, underscoring the powerful potential of foundation model in capturing morphological features for biological discovery. Collectively, SpaFoundation open new avenues for image-guided spatial omics and holds potential to advance histology-driven computational analyses in spatial biology.

Methods

Pretraining Dataset for SpaFoundation

We adopt the HEST-1K spatial transcriptomics dataset, curated by Guillaume et al., as the pretraining dataset for SpaFoundation. HEST-1K encompasses spatial transcriptomics data generated using three distinct technologies, i.e., Spatial Transcriptomics, Xenium, and Visium (including Visium HD), with each sample paired with a corresponding histological image. The dataset includes 1,229 samples from 153 public or internal cohorts, spanning two species: Homo sapiens (47.2%) and Mus musculus (52.8%). After tissue detection and segmentation, HEST-1K yields approximately 2.1 million expression–morphology pairs (each patch sized 224×224 pixels) and 76.4 million nuclei. Notably, three datasets used in the downstream tasks, i.e., HER2+, cSCC, and HBCIS, are excluded from pre-training. The remaining subset, referred to as HEST-1K-3, includes 1,113 samples across 26 distinct tissue types.

SpaFoundation model architecture

SpaFoundation is pretrained using the iBOT[36] framework to learn generalizable pathological representations useful for spatial transcriptomics analysis (Fig. 1a). To this end, we employ a ViT-B/16-based teacher-student network that learns patch-level representations by encoding paired local and global crops, as well as masked global crops. The model is optimized through two complementary objectives, i.e., self-distillation and masked image modeling (MIM). The self-distillation component minimizes cross-entropy loss between the [CLS] tokens of different crops from the same pathological image, capturing high-level semantic information. The MIM branch reconstructs the patch tokens learned from the original and masked global crops, enabling the model to preserve fine-grained local structure features. By jointly optimizing the self-distillation and MIM objectives, SpaFoundation facilitates comprehensive modeling of multi-level semantics in pathological images, resulting in biologically meaningful and robust feature representations for various downstream tasks. Next, we will elaborate on the components in detail.

Self-distillation for learning high-level semantic features. Tissue structures often exhibit significant variability in shape, size, and orientation in the pathological images, posing significant challenges for learning representations that can be generalizable for diverse downstream analysis. To address these challenges and enhance semantic feature learning, we incorporate a self-distillation mechanism into the SpaFoundation model backbone. This mechanism is designed to capture high-level semantic features by encouraging consistency across multiple views of the same image. Specifically, given an input image patch, we generate m global crops (with crop ratio ranging from 32% to 100%) and n local crops (with crop ratio ranging from 5% to 32%) through random cropping and image augmentation operations. Although the local and global crops are derived from the same image, they capture different structure features due to variations in position and scale. To further increase the learning robustness, we create masked global crops by dividing each global crop into non-overlapping image patches and randomly masking some

percentage of the image patches.

With the crops from different views as inputs, the self- distillation mechanism learns patch-level representations by training a student network to match the output of a given teacher network. Specifically, the teacher network encodes the global crops and outputs global representations, while the student network encodes both the local and masked global crops, and outputs local and masked global representations. For each crop view, its output representation matrix comprises a [CLS] token and patch tokens. Notably, the [CLS] token represents the semantic representation of the crop view. By minimizing the cross-entropy loss between these [CLS] tokens, SpaFoundation is encouraged to learn high-level semantic features. The self-distillation loss function $\mathcal{L}_{\mathrm{SD}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{SD}} = -P_{\theta_s}(X_s)^T log P_{\theta_t}(X_t)$$

where $\mathrm{P}(\cdot)$ denotes the probability distributions predicted by the teacher-student network. X_t and X_s represent the [CLS] tokens from the teacher and student networks, respectively. θ_{t} and θ_{s} refer to the parameters of the teacher and student networks, respectively.

Masked image modeling for learning local semantic features. Building on BEiT[37], we introduce masked image modeling (MIM) to enhance representation learning by predicting the visual tokens that are corresponding to the masked patches. This component enables the model to capture local semantic features. Specifically, the teacher network takes as input the original global crops and outputs original patch tokens. The student network takes as inputs the masked global crops and outputs predicted patch tokens. The pre-training objective of MIM is to minimize the cross-entropy loss between the original and predicted patch tokens, guiding the student network to recover masked image information and thereby enhancing its ability to capture local semantic features. Formally, the MIM loss function $\mathcal{L}_{\mathrm{MIM}}$ is defined as follows:

$$\mathcal{L}_{\mathrm{MIM}} = -\sum_{j=1}^{N} m_j \cdot P_t\left(h_j^t\right)^T log P_s(h_j^s),$$

where N is the number of patches and $\mathrm{m} \in [0,1]$ denotes binary mask vector, with $m_j = 1$ indicating that the j-th patch of the image is masked, h_j^t and h_j^s denote the j-th patch token from the teacher and student networks, respectively. $\mathrm{P}(\cdot)$ denotes the probability distribution from the tokenizer.

Model optimization. The pretraining model is optimized jointly by the self-distillation loss $\mathcal{L}_{SD}$ and the MIM loss $\mathcal{L}_{\mathrm{MIM}}$. The overall loss is defined as follows:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{SD} + \lambda_2 \mathcal{L}_{\mathrm{MIM}},$$

where λ_1 and λ_2 are weighting factors measuring the contributions of the self-distillation and MIM losses. During pretraining, the parameters of the teacher network are exponentially moving averaged (EMA) to the parameters of the teacher network. This updating strategy helps enhance model stability.

SpaFoundation pretraining details

In this study, the SpaFoundation model was trained for a total of 10 epochs. SpaFoundation was trained using a distributed training strategy across 8 Tesla V100 NVlink GPUs (each with 32GB of memory), enabling parallel computation with a batch size of 64 per GPU. The model architecture was based on Vision Transformer (ViT), specifically the vit_base variant. We employed the AdamW optimizer, initialized with a learning rate of 0.0002 and dynamically adjusted using a cosine annealing scheduler. To improve training efficiency and stability, we implemented mixed precision training, gradient clipping, and initially frozen the final layer to facilitate better convergence. For data augmentation, we generated two global views and ten local views for each input image patch. Additionally, standard image preprocessing, including random cropping, color jitter, and horizontal flipping, were applied.

Evaluation

In this study, we evaluated SpaFoundation's performance across four tasks, including spatial gene expression prediction, high-resolution gene expression inference, tumor region detection, and spatial clustering. To better assess model performance, we used evaluation metrics tailored to each task. Specifically, both spatial gene expression prediction and high-resolution gene expression inference tasks were evaluated using the Pearson correlation coefficient (PCC), which measures the linear correlation between raw and predicted gene expression values. For tumor region detection, accuracy (ACC) and F1 score were employed, along with receiver operating characteristic (ROC) curves and the area under the ROC curve (AUC). In the spatial clustering, the Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) were used to evaluate the agreement between known and predicted labels. These evaluation metrics allowed for a comprehensive analysis and comparison of model capability across different tasks.

Spatial gene expression prediction task. We first evaluated the performance of SpaFoundation in predicting spatial gene expression profiles using two spatial transcriptomics datasets, i.e., cSCC [29] and HER2+[30]. The cSCC cohort comprises 12 tissue sections from 4 patients, while the HER2+ cohort includes 36 sections from 8 patients. To evaluate SpaFoundation's robustness, we designed two prediction scenarios, i.e., highly expressed genes (HEG) expression prediction and highly variable genes (HVG) expression prediction. For the HEG prediction, we estimated the expression levels of the top 250

highly expressed genes, and for the HVG prediction, we predicted the expression levels of the top 1,000 most variable genes across all the sections.

To better imitate real-world scenarios, we adopted a leave-one-patient-out cross-validation strategy, in which tissue sections from a held-out patient were used as the test set, while the remaining sections served as the training set. The model was fine-tuned on the training set using the pretrained weights of SpaFoundation as initialization. Gene expression in the test set was predicted using a lightweight regressor composed of a layer normalization followed by a linear layer. The model was optimized using mean square error (MSE) loss.

High-resolution spatial gene expression prediction task. We further examined the performance of SpaFoundation in inferring high-resolution spatial gene expression from low-resolution spatial transcriptomics data. We employed a breast cancer Visium HD dataset, which profiles tissue sections at two spatial resolutions, i.e., 8 µm, and 16 µm. We fine-tuned the pre-trained SpaFoundation model on the 16 µm scaling resolution data (~140,000 spots), and then predicted gene expressions at the 8 µm resolution (~580,000 spots). Notably, we predicted the expression levels of the top 1,000 most variable genes, as well as two breast cancer marker genes, *ERBB2* and *CD24*. Image features were extracted using SpaFoundation and fed into the same lightweight regressor (as described in the previous task) to predict gene expression. The model was trained using MSE loss, without applying any image augmentation. We compared SpaFoundation with ST-Net, a state-of-the-art method capable of performing this task. Other baseline methods such as HistoGene, His2ST, BLEEP, TRIPLEX, and THItoGene were excluded due to their high computational complexity.

Tumor region detection task. We next assessed the effectiveness of SpaFoundation in detecting tumor regions in the cancer tissue samples using the HBCIS dataset[17], which includes 68 tissue sections from 23 patients. Five patients were selected as the test cohort, while the remaining patients were used for training. The test patients were specifically chosen to ensure that each molecular subtype was represented by at least one patient.

To detect tumor regions, we fine-tuned the SpaFoundation model on the training set to extract generalizable pathological representations. These representations were then input into a classifier to predict whether a given patch belonged to a tumor region. The classifier comprises two sequential modules, each composed of a layer normalization followed by a linear transformation. The model was trained in a supervised manner using binary cross-entropy loss. For assessing SpaFoundation's ability in extracting meaningful pathological representations, we compared it with two popular baseline models, i.e., ResNet_in1k (ResNet-50)[31] and VGG_in1k (VGG-16)[32]. Both were initialized with weights pre-trained on the ImageNet-1K dataset.

Spatial clustering task. We also examined the capability of SpaFoundation on spatial clustering task. For this task, we used the same HER2+ dataset as the spatial gene expression prediction task. This dataset includes 36 tissue sections from 8 patients. For each patient, one section was manually annotated by expert pathologists, resulting in 8 annotated sections (i.e., A1, B1, C1, D1, E1, F1, G2, H1). The spatial clustering task was designed as a zero-shot setting. Notably, taking as inputs tissue image patches, we extracted patch-level representations using the pre-trained SpaFoundation model without any fine-tuning. These representations were then subjected to dimensionality reduction via PCA, followed by unsupervised clustering using the Leiden algorithm[38].

SpaFoundation fine-tuning details

SpaFoundation was fine-tuned for 20 (25 for HER2+), 20, 1 epochs for the spatial gene expression prediction, high-resolution spatial gene expression prediction, tumor region detection tasks, respectively. The spatial clustering task was evaluated in a zero-shot setting and thus no fine-tuning was performed.

The spatial gene expression prediction and spatial clustering tasks were executed on an NVIDIA RTX A6000 GPU, while the high-resolution prediction and tumor detection tasks were run on an NVIDIA RTX 4090 GPU. All models were trained with a batch size of 64 using the Adam optimizer with an initial learning rate of 1×10^{-5}. For the gene expression prediction tasks, a StepLR learning rate scheduler was applied, reducing the learning rate by a factor of 0.9 every 50 epochs.

To enhance robustness to orientation and spatial variation, online data augmentation was applied during training. Each image was subjected to random horizontal and vertical flips with a 50% probability, as well as random rotations selected from {−90°, 0°, 90°, 180°}. For automated resolution search, the resolution was scanned from 0.1 to 3.0 in steps of 0.01. The number of neighbors for k-nearest neighbor (KNN) graph construction was fixed at 6.

Baseline methods

To validate the effectiveness of SpaFoundation, we compared SpaFoundation with HistoGene[20], His2ST[19], THItoGene[21], BLEEP[22], and TRIPLEX[23] in the spatial gene expression prediction task. For high-resolution gene expression prediction, SpaFoundation was compared with ST-Net[17]. To evaluate the ability of SpaFoundation in learning biologically meaningful representations, we compared SpaFoundation with RestNet[31] and VGG[32] for tumor region detection. To validate the performance of SpaFoundation in the spatial clustering, we compared SpaFoundation with state-of-the-art methods, including Scanpy[33], SpaGCN[34], and stLearn[35].

HistoGene. HistoGene is a vision transformer model developed for spatial

gene expression prediction using both pathological image and spatial coordinates as input. We followed the original parameter settings reported in the published paper. The batch size and number of training epoch were set to 1 and 100, respectively. To ensure consistency, we applied the same data augmentation strategy used in SpaFoundation when processing training images.

His2ST. His2ST is an extended version of HistoGene that integrates a graph neural network with a zero-inflated negative binomial distribution to predict gene expression from histological image. We followed the original configuration as described in the paper. In particular, the batch size and number of training epochs were set to 1 and 350, respectively. Consistent with the HistoGene implementation, we applied data augmentation strategy and used deterministic sample sorting during data preprocessing.

THItoGene. THItoGene is a deep learning framework that integrates dynamic convolution and capsule networks to adaptively capture latent molecular signals from histological images. We followed the default parameter settings, with the batch size and number of training epochs set to 1 and 200, respectively. To ensure reproducibility, we applied training data augmentation strategy, implemented deterministic sample sorting, and fixed the random seed.

BLEEP. BLEEP is a bimodal embedding model that predicts spatial gene expression from H&E-stained histology images using contrastive learning. During inference, BLEEP provides three strategies. We adopted the simple averaging method with $K = 50$, which demonstrated the best performance in the original study. The model was trained using two GPUs with a batch size of 64. All other settings followed the original configuration reported in the BLEEP paper, including training for 150 epochs and using ResNet-50 as the image encoder.

TRIPLEX. TRIPLEX is a deep learning framework designed to predict spatial gene expression from Whole Slide Images (WSIs). As the model provides pre-extracted image features, we did not apply online image augmentation during inference. To ensure consistency with other baseline models, we disabled two additional configurations originally used in TRIPLEX: test set validation and expression smoothing.

ST-Net. ST-Net is a machine learning model for predicting spatial transcriptomics measurements from haematoxylin-and-eosin-stained pathology slides. It employs a DenseNet-121 model, pre-trained on ImageNet, to extract tissue features and predict spatial gene expression. We followed the default settings described in the original paper and used the ST-Net implementation available in the TRIPLEX codebase.

ResNet & **VGG.** ResNet and VGG are classic and widely adopted convolutional

neural networks in computer vision. In this study, we employed ResNet-50 and VGG-16, both pre-trained on the ImageNet-1K dataset, for the tumor region detection task. These models were used to extract image features and perform binary classification of tumor versus non-tumor regions. To ensure consistency with SpaFoundation, we adopted the same training settings, including data preprocessing, augmentation strategies, loss functions, and optimization protocols.

Scanpy. Scanpy is a widely used toolkit for single-cell transcriptomic analysis. In this study, we included Scanpy as a baseline method in the spatial clustering task.

SpaGCN. SpaGCN is a graph-based method that integrates gene expression, spatial coordinates, and histological image features to identify spatial domains. We compared SpaFoundation with spaGCN in the spatial clustering task. We followed the default parameter to run the model.

stLearn is a spatial clustering method that integrates gene expression and spatial information for spatial domain identification. We compared SpaFoundation with stLearn in the spatial clustering task, and adopted the default parameters to run the model.

Data availability

We analyzed the following datasets used in SpaFoundation, including the pretraining dataset and those employed for evaluating downstream task performance:(1) a dataset for ST and histology image analysis dataset HEST-1K(https://huggingface.co/datasets/MahmoodLab/hest), (2) human HER2-positive breast cancer ST data reported in Anderson et al. (https://github.com/almaan/her2st/), (3) Human Squamous Cell Carcinoma ST data reported in Andrew et al. (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE144240), (4) Human breast cancer ST data reported in He et al. (https://data.mendeley.com/datasets/29ntw7sh4r/5), and (5) 10×Visium HD Human Breast Cancer data (https://www.10xgenomics.com/cn/datasets/visium-hd-cytassist-gene-expression-libraries-human-breast-cancer-ff-ultima).

Code availability

The codes were implemented in Python and have been released on GitHub (https://github.com/NingZhangCSUBio/SpaFoundation).

Acknowledgements

This work is supported in part by the National Natural Science Foundation of China under Grant (No.62225209 to M.L.) and the High Performance Computing Center of Central South University.

Author contributions

M.L. and Y.L. conceptualized and supervised the project. N.Z., Y.L., and S.X. designed the model. N.Z. developed the SpaFoundation software. N.Z., Y.L., and M.L. wrote the manuscript. N.Z., Y.L., S.X., Y.J., P.J., S.C. and J.C. performed the data analysis. N.Z., Y.L., and M.L. contributed to figure design and generation.

Competing interests

The authors declare no competing interests.

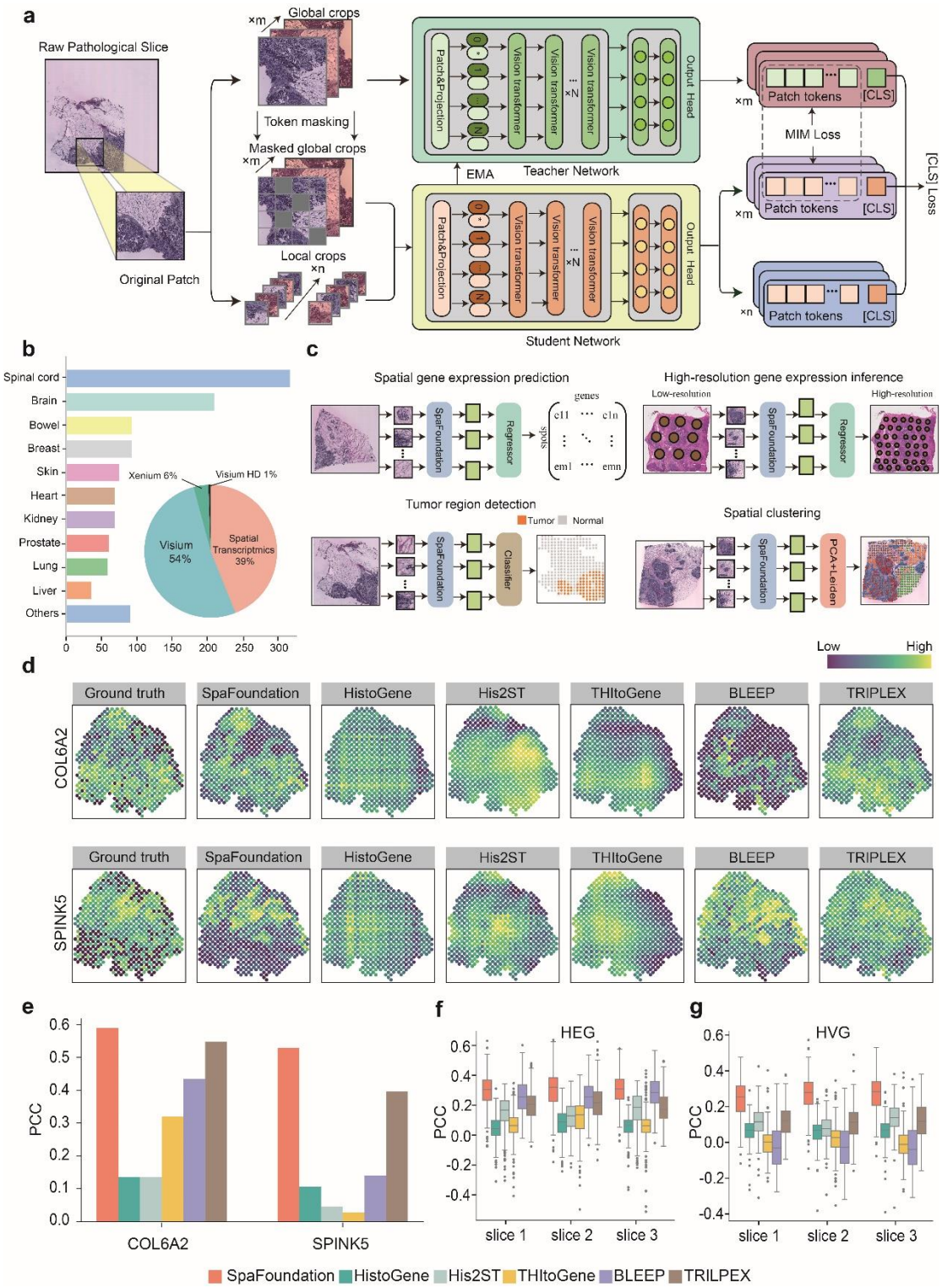

Fig.1 Overview of SpaFoundation and superior performance in spatial gene expression prediction. a, The workflow of pretraining SpaFoundation using pathological images within the iBOT framework, which is based on a teacher-student architecture and incorporates self-distillation and masked image modeling (MIM) components. **b,** Statistics of the HEST-1K-3 pre-training dataset, comprising 1,113 spatial transcriptomics whole-slide images (WSIs) from 26 organ types generated using three distinct technologies. **c,** Illustration of four downstream analysis tasks, including spatial gene expression prediction,

high-resolution gene expression inference, tumor region detection, and spatial clustering. **d-g**, Results on the cSCC dataset for two prediction tasks: high expression genes (HEG) and highly variable genes (HVG). **d,** Ground truth and spatial distributions of genes *COL6A2* and *SPINK5* predicted by SpaFoundaiton and baseline methods, including HistoGene, His2ST, THItoGene, BLEEP, and TRIPLEX. **e**, Quantitative evaluation of different methods based on Pearson correlation coefficients (PCC) for genes *COL6A2* and *SPINK5*. **f**, Boxplots of the PCC metric for the HEG prediction. In the box plot, the center line denotes the median, box limits denote the upper and lower quartiles, and whiskers denote the 1.5 times the interquartile range. n = 250 genes. **g**, Boxplots of the PCC metric for the HVG prediction. n = 171 genes.

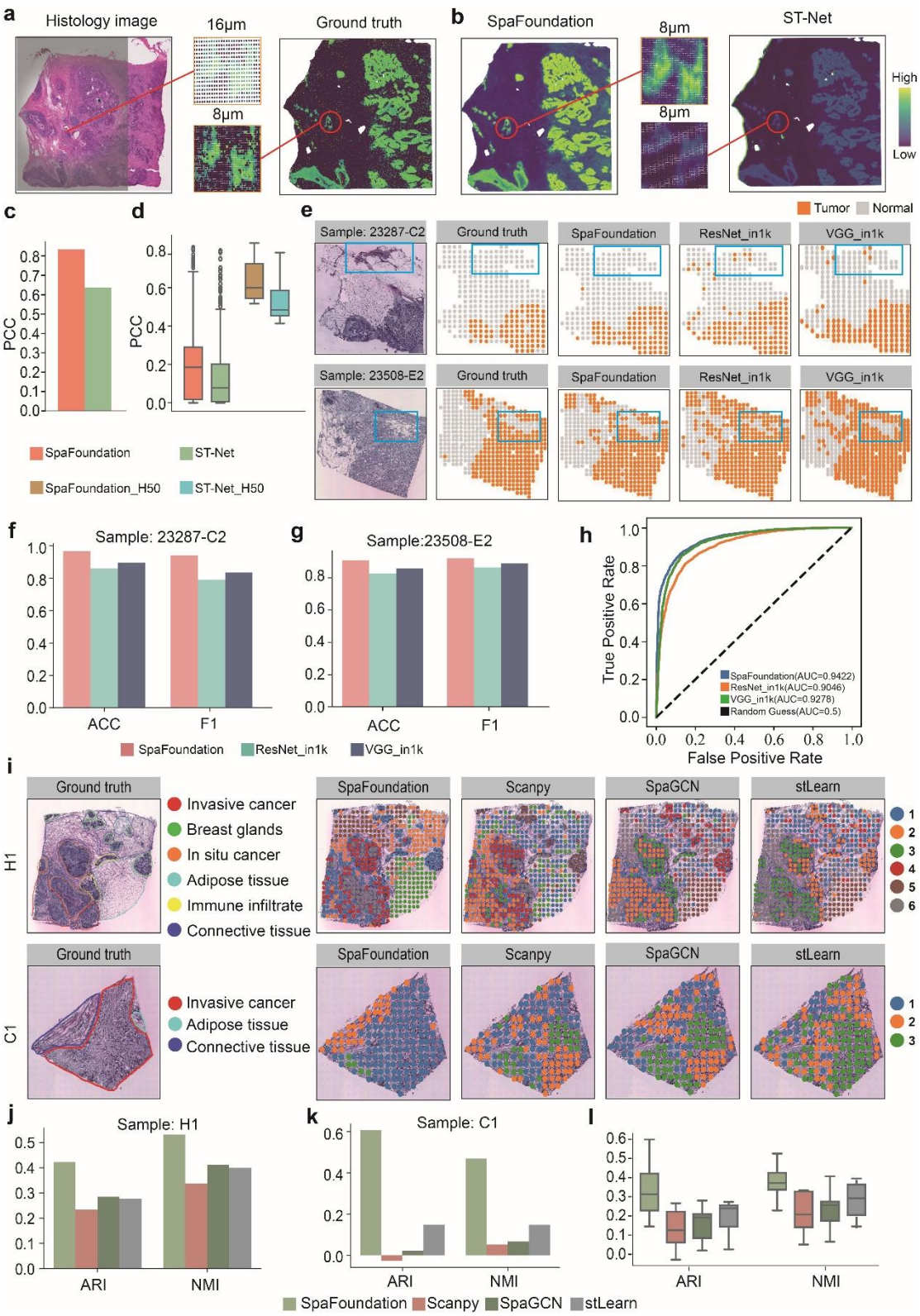

Fig.2 SpaFoundation accurately infers high-resolution gene expression and exhibits robust performance in tumor detection and spatial clustering. a-d, Results on the Visium HD Breast Cancer dataset for the High-resolution gene expression inference task. **a,** Histology image and ground truth expression of *ERBB2* at different resolutions. **b**, Comparison of spatial distributions of gene *ERBB2* predicted by SpaFoundation and ST-

Net. **c,** Quantitative evaluation of different methods based on PCC for genes *ERBB2*. **d,** Boxplots of the PCC metric for the high-resolution gene expression prediction. n = 1002 genes. **e-h,** Results on the HBCIS dataset for the tumor region detection task. **e,** Histology images, ground truth, and spatial visualization of tumor regions predicted by SpaFoundation and baseline models, including ResNet_in1k and VGG_in1k, for the 23287-C2 and 23508-E2 samples. **f,** Quantitative evaluation of different methods for the 23287-C2 sample using ACC and F1 score. **g**, Quantitative evaluation of different methods for the 23508-E2 sample. **h**, Receiver operating characteristic (ROC) curves for tumor region detection across all test samples. **i-l,** Performance of SpaFoundation in spatial clustering on the the HER2+ dataset. **i**, Ground truth and spatial distributions of clustering results predicted by SpaFoundation and baseline methods, including Scanpy, SpaGCN, and stLearn. **j**, Quantitative evaluation of different methods for the H1 sample using ARI and NMI. **k**, Quantitative evaluation of different methods for the C1 sample. **l**, Boxplots of ARI and NMI for the spatial clustering. n=8 samples.

Reference

1. Yayon, N. *et al.* A spatial human thymus cell atlas mapped to a continuous tissue axis. *Nature* **635**, 708–718 (2024).
2. Vannan, A. *et al.* Spatial transcriptomics identifies molecular niche dysregulation associated with distal lung remodeling in pulmonary fibrosis. *Nat. Genet.* **57**, 647–658 (2025).
3. Polonsky, M. *et al.* Spatial transcriptomics defines injury specific microenvironments and cellular interactions in kidney regeneration and disease. *Nat. Commun.* **15**, 7010 (2024).
4. Li, L. *et al.* An organ-wide spatiotemporal transcriptomic and cellular atlas of the regenerating zebrafish heart. *Nat. Commun.* **16**, 3716 (2025).
5. Qian, X. *et al.* Spatial transcriptomics reveals human cortical layer and area specification. *Nature* 1–11 (2025) doi:10.1038/s41586-025-09010-1.
6. Yang, W. *et al.* Deciphering cell–cell communication at single-cell resolution for spatial transcriptomics with subgraph-based graph attention network. *Nat. Commun.* **15**, 7101

(2024).

7. Zhang, J. *et al.* Spatiotemporally resolved transcriptomics reveals the cellular dynamics of human retinal development. *Nat. Commun.* **16**, 2307 (2025).
8. Feng, J. *et al.* High-resolution spatial transcriptomics and cell lineage analysis reveal spatiotemporal cell fate determination during craniofacial development. *Nat. Commun.* **16**, 4396 (2025).
9. Three-dimensional intact-tissue sequencing of single-cell transcriptional states | Science. https://www.science.org/doi/abs/10.1126/science.aat5691.
10. Highly sensitive spatial transcriptomics at near-cellular resolution with Slide-seqV2 | Nature Biotechnology. https://www.nature.com/articles/s41587-020-0739-1.
11. Chen, A. *et al.* Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777-1792.e21 (2022).
12. Integrated spatial genomics reveals global architecture of single nuclei | Nature. https://www.nature.com/articles/s41586-020-03126-2.
13. Liu, J. *et al.* Concordance of MERFISH spatial transcriptomics with bulk and single-cell RNA sequencing. *Life Sci. Alliance* **6**, (2023).
14. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging | Nature Biotechnology. https://www.nature.com/articles/s41587-022-01483-z.
15. Zhang, D. *et al.* Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat. Biotechnol.* **42**, 1372–1377 (2024).
16. Bergenstråhle, L. *et al.* Super-resolved spatial transcriptomics by deep data fusion.

Nat. Biotechnol. **40**, 476–479 (2022).

17. He, B. *et al.* Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
18. Li, S., Gai, K., Dong, K., Zhang, Y. & Zhang, S. High-density generation of spatial transcriptomics with STAGE. *Nucleic Acids Res.* **52**, 4843–4856 (2024).
19. Zeng, Y. *et al.* Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Brief. Bioinform.* **23**, bbac297 (2022).
20. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. Preprint at https://doi.org/10.1101/2021.11.28.470212 (2021).
21. Jia, Y., Liu, J., Chen, L., Zhao, T. & Wang, Y. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. *Brief. Bioinform.* **25**, bbad464 (2023).
22. Xie, R. *et al.* Spatially Resolved Gene Expression Prediction from H&E Histology Images via Bi-modal Contrastive Learning. *NeurIPS* (2023).
23. Chung, Y., Ha, J. H., Im, K. C. & Lee, J. S. Accurate Spatial Gene Expression Prediction by integrating Multi-resolution features. Preprint at http://arxiv.org/abs/2403.07592 (2024).
24. Lecun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
25. Vaswani, A. *et al.* Attention is All you Need. in *Advances in Neural Information Processing Systems* vol. 30 (Curran Associates, Inc., 2017).

26. Chen, R. J. *et al.* Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).

27. Lu, M. Y. *et al.* A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).

28. Wang, X. *et al.* A pathology foundation model for cancer diagnosis and prognosis prediction. *Nature* **634**, 970–978 (2024).

29. Multimodal Analysis of Composition and Spatial Architecture in Human Squamous Cell Carcinoma: Cell. https://www.cell.com/cell/fulltext/S0092-8674(20)30672-3?uuid=uuid%3A17a76bad-26cc-4068-920b-aa7ed206a71e.

30. Spatial deconvolution of HER2-positive breast cancer delineates tumor-associated cell type interactions | Nature Communications. https://www.nature.com/articles/s41467-021-26271-2.

31. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition. in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 770–778 (2016). doi:10.1109/CVPR.2016.90.

32. Simonyan, K. & Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. Preprint at https://doi.org/10.48550/arXiv.1409.1556 (2015).

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

34. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network | Nature Methods. https://www.nature.com/articles/s41592-021-01255-8.

35. Pham, D. T., Tan, X., Xu, J., Grice, L. F. & Nguyen, Q. H. stLearn: integrating spatial location, tissue morphology and gene expression to find cell types, cell-cell interactions and spatial trajectories within undissociated tissues. *bioRxiv* (2020) doi:10.1101/2020.05.31.125658.

36. Zhou, J. *et al.* iBOT: Image BERT Pre-Training with Online Tokenizer. Preprint at http://arxiv.org/abs/2111.07832 (2022).

37. Bao, H., Dong, L., Piao, S. & Wei, F. BEiT: BERT Pre-Training of Image Transformers. Preprint at https://doi.org/10.48550/arXiv.2106.08254 (2022).

38. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).