nature communications

Review article

https://doi.org/10.1038/s41467-025-58989-8

Combining spatial transcriptomics with tissue morphology

Received: 26 August 2024

Accepted: 4 April 2025

Published online: 13 May 2025

Check for updates

Eduard Chelebian, Christophe Avenel & Carolina Wählby

Spatial transcriptomics has transformed our understanding of tissue architecture by preserving the spatial context of gene expression patterns. Simultaneously, advances in imaging AI have enabled extraction of morphological features describing the tissue. This review introduces a framework for categorizing methods that combine spatial transcriptomics with tissue morphology, focusing on either translating or integrating morphological features into spatial transcriptomics. Translation involves using morphology to predict gene expression, creating super-resolution maps or inferring genetic information from H&E-stained samples. Integration enriches spatial transcriptomics by identifying morphological features that complement gene expression. We also explore learning strategies and future directions for this emerging field.

Before the advent of spatial transcriptomics, spatial information would typically be lost as single-cell omics relied on tissue dissociation[1–4]. Preserving the spatial context of cells provides a more comprehensive view of cellular heterogeneity, interaction, and tissue architecture, leading to deeper insights into the molecular landscape of disease[5–7]. The early analysis methods in spatial transcriptomics naturally evolved from established analysis techniques developed by the single-cell community[8]. However, this reliance often led to an oversight of the spatial and morphological relationships present within tissues.

In parallel, developments in artificial intelligence and machine learning have significantly enhanced our ability to analyze and interpret image data. Deep learning algorithms, convolutional neural networks (CNNs), and image segmentation techniques have enabled more precise analysis of biological tissues[9,10]. Spatial transcriptomics, which captures spatially-resolved data and usually includes imaging data, naturally intersects with such technologies. By combining the gene expression information from spatial transcriptomics and morphological features from imaging, we can achieve a more holistic understanding of tissue architecture[11]. Nonetheless, there are some challenges due in part to the disconnect between the expertise of both communities. It is inherently hard to simultaneously utilize two different data sources of such high dimensionality as spatial transcriptomics and imaging data. While obtaining features from spatial transcriptomics is relatively established and interpretable, the patterns often directly represent local gene expression, and extracting relevant morphological features from images requires more understanding and careful validation.

Combining advancements in imaging AI into spatial transcriptomics would benefit from a structured approach to leverage these combined data sources effectively. Previous reviews and benchmarks have explored separate aspects of the combination[12,13] or, more generally, the applications of AI for spatial transcriptomics[14,15]. This Review introduces a comprehensive framework to understand such combination methods by analyzing the distinct ways morphological features can be utilized. Image morphology can either be *translated* into features that correlate with spatial transcriptomics, taking advantage of the ease and cost-effectiveness of acquiring it, or *integrated* with spatial transcriptomics to provide a richer description of the sample. These approaches are inherently conflicting: translation focuses on gene-correlated features, while integration searches for complementary information for a fuller understanding. This Review aims to clarify these distinctions to optimize the use of morphological and molecular data together.

Translation-integration framework for joint analysis of morphology and spatial transcriptomics

Spatial transcriptomics, whether imaging-based or sequencing-based, in practice yields a grid of spatial positions together with gene expression information[2]. The process for obtaining the gene expression is quite standardized for the different modalities of spatial transcriptomics. What users receive from sequencing-based solutions such

Department of Information Technology and SciLifeLab, Uppsala University, Uppsala, Sweden. e-mail: carolina.wahlby@it.uu.se

as 10X Visium, slide-seq[16], or stereo-seq[17] is a matrix of spots or cells representing gene expression, together with their spatial coordinates, ready for downstream analysis. Similarly, imaging-based solutions, regardless if they are in situ hybridization-based[18,19], or in situ sequencing-based[20,21], yield a stack of images with signals that need to be detected and decoded to retrieve the location of specific genes. Even if the specific methods to do so can vary, the general workflow is usually the same. The individual detections can then be aggregated according to a fixed pattern or, e.g., per cell to, again, construct a matrix of gene expression together with spatial coordinates. The gene expression from spatial transcriptomics approaches on their own has shown their biological relevance in countless works[5–7]. But this information is fixed and, once the appropriate analyses are in place, it is not possible to tweak the gene expression to obtain more relevant information.

Tissue morphology, typically enhanced by stains such as hematoxylin-eosin (H&E) or DAPI prior to imaging, is often paired with spatial transcriptomics and introduces an additional layer of information. In order to combine the gene expression with the image information, it is common to pair them by extracting patches using the gene expression coordinates as patch centers, as shown in Fig. 1a. The process of obtaining meaningful characteristics from images is known as feature extraction. Traditionally, features were extracted using hand-crafted algorithms with known specific outcomes in mind. However, learning-based algorithms have proven to capture more powerful representations in a variety of medical tasks[9,10]. Unlike gene expression from spatial transcriptomics, the features from images can be learned to be task-specific. For instance, the morphological features extracted for performing cancer grading can be very different from the morphological features for detecting mitosis for the same piece of tissue[22]. Thus, the framework that we propose focuses on the design choices regarding the learned morphological features, their *relevance*, and their *shared information* with gene expression.

The features we extract from morphological images should clearly be as relevant as possible, which is in itself a noteworthy task. General task-independent irrelevant information in a tissue slide could be artifacts from staining and imaging. We consider relevance as the importance of features for the specific task, which can differ across contexts. Further, relevance can usually only be measured in these cases by assessing the performance of the features in a downstream task. However, relevance in itself is not sufficient for a joint analysis of morphology and spatial transcriptomics. The morphological features need to have specific synergies with gene expression patterns in order to consider them appropriate for different joint analyses. For example, some features may contain relevant information solely due to their similarity with gene expression, while others might be relevant independently of gene expression.

If we construct a plot with these two dimensions, relevant information in the morphology features on the *y*-axis and shared information between morphology and gene expression on the *x*-axis, we obtain the four quadrants depicted in Fig. 1b. The quadrants I and II (translation and integration) represent scenarios in which features can be used for joint analysis with spatial transcriptomics, while the quadrants II and IV (noise and overestimation) represent scenarios in which features should not be used for joints analyses with spatial transcriptomics:

I. *Translation*: in this scenario, the morphological features contain a high amount of relevant information and share a high amount of information with gene expression. These features include the same information as the task-relevant part of the gene expression. This scenario is ideal for gene expression prediction to, for instance, generate super-resolution maps or infer genetic information from clinical H&E-stained samples without incurring in the costs. These features are typically obtained by specifically training deep learning models for that purpose.
II. *Integration*: in this scenario, the morphological features contain a high amount of relevant information but do not share information with gene expression. These features contain relevant information that has not been captured by the gene expression. This scenario is ideal for spatial domain identification due to the time uncoupling between changes in gene expression and its effect on morphology. These features need to be more general as they should complement the information in the gene expression and not be redundant.
III. *Noise*: in this scenario, the morphological features do not contain relevant information, nor do they share information with gene expression. These features capture non-relevant variations, such as sample-to-sample or staining variations, which cannot be leveraged in a joint analysis with spatial transcriptomics, as they will confuse the analysis.
IV. *Overestimation*: in this scenario, the morphological features do not contain relevant information but share information with gene expression. These features cover the gene expression that is not task-relevant and thus can lead to an overestimation of the predictive power in, for instance, translation tasks. These can be housekeeping or constitutive genes that may not be relevant for the specific downstream task.

This framework already enables rethinking the way we design the morphological features (e.g., by training a neural network) for the different tasks and whether the joint analysis is even valuable or not. For instance, if we have morphological features that are highly correlated with gene expression, integration is meaningless, if not harmful, as we would be including redundant information in our model. Equally, doing translation with morphological features that do not share information with gene expression would not be effective.

The formal definition of this framework is presented in Box 1, and we present a practical example in Box 2.

Morphology translation for gene expression prediction

Morphology translation involves identifying morphological features that spatially correlate with gene expression patterns. The primary application in this scenario is gene expression prediction. The goal is to learn morphological features that are highly correlated with gene expression so that these features can be used independently of gene expression data, either on new samples (e.g., parallel tissue slices in attempts to re-create 3D volumes) or in regions of a sample where spatial gene expression data is unavailable or sparse (see Fig. 2).

Note that, within our framework, we specifically refer to the relationship of morphological images with spatial transcriptomics. Previous work has also tried to predict spatial gene expression using non-spatial sources[23–26].

Using spatial gene expression data, the most common application has been to try and predict 10X Visium and other sequencing-based spatial transcriptomics expression from H&E images, and we will start this section with a historical perspective of such methods.

To our knowledge, this field began to take shape with the development of ST-Net[27], paradoxically entitled "*Integrating* spatial gene expression and breast tumor morphology via deep learning," which, per our definition, was a translation application rather than integration. The authors identified known breast cancer biomarkers and, even though the correlation values were moderate, they showed that the translation was relevant.

The still-unpublished HisToGene[28] is one of the first works available introducing the concept of super-resolution for spatial transcriptomics. The authors translate the gene expression to a dense map of the whole image, theoretically being able to achieve pixel-level expression. Monjo et al. trained DeepSpaCE[29] to achieve super-resolution gene expression of areas between the spots. Additionally,

https://doi.org/10.1038/s41467-025-58989-8

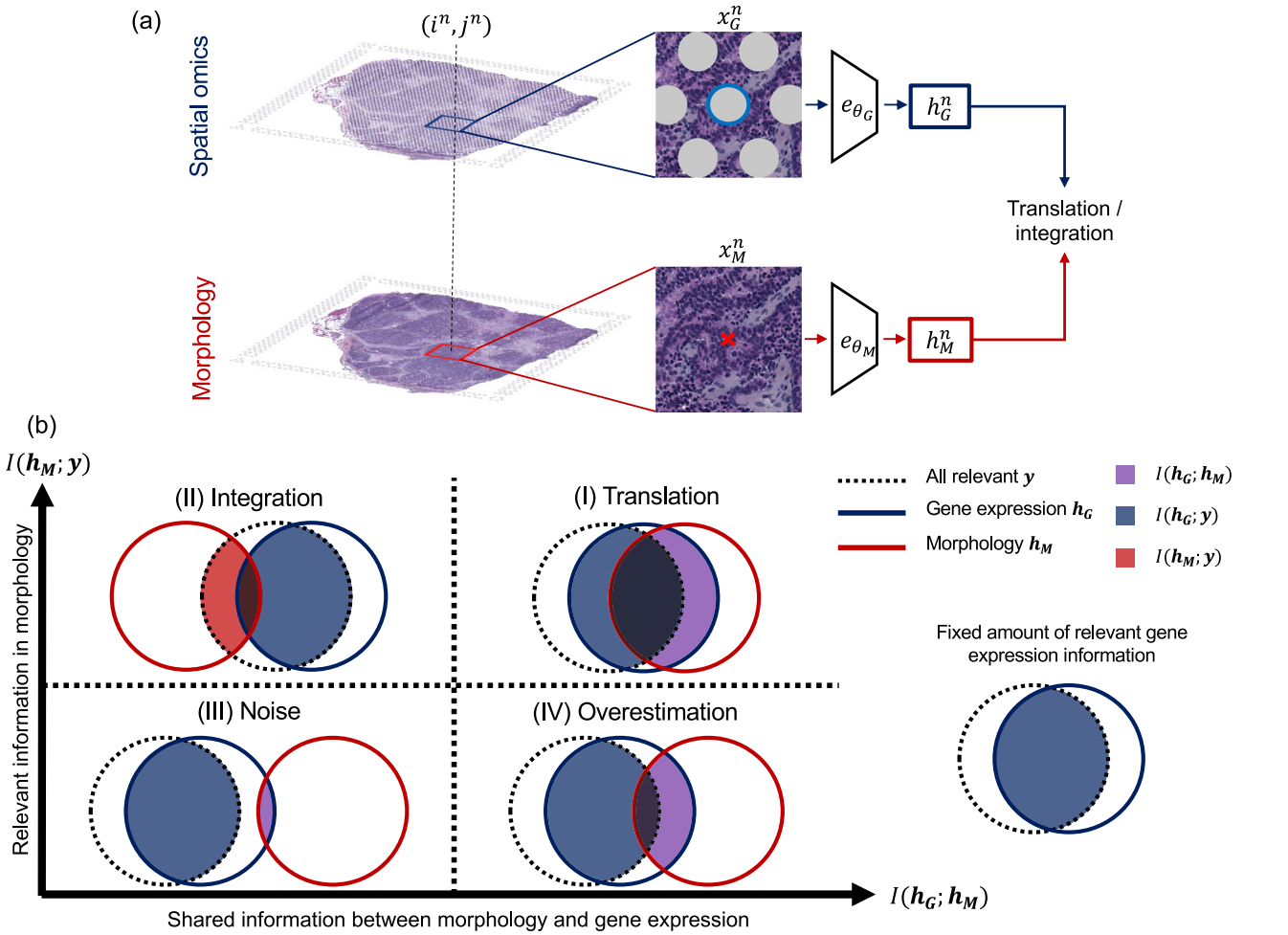


Fig. 1 | Translation-integration framework. a Feature extraction from spatial transcriptomics and morphology on 10X Visium spatial transcriptomics and H&E[57]. We use the nth spot in Visium x_G^n as the center to extract the nth image patch x_M^n. The modality-specific encoders e_{θ_G} and e_{θ_M} will output the gene expression feature vector h_G^n and the morphological feature vector h_M^n for position (i^n, j^n). **b** Intuition of the framework. By assuming the amount of relevant spatial gene expression information is fixed (solid blue area), we get four different scenarios depending on the morphological features. We express these four scenarios by the quadrants formed when presenting the relevance of the morphological features as the y-axis and their shared information with the gene expression/spatial transcriptomics features as the x-axis.

they performed translation on consecutive sections in a semi-supervised way.

STimage[30] introduced a probabilistic framework using CNNs and negative binomial regression, with a log-likelihood loss to estimate gene expression distributions, enhancing robustness and interpretability. Similarly, Hist2ST[31] combined convolutional mixers, Transformers, and graph neural networks (GNNs) while employing a zero-inflated negative binomial loss, integrating spatial dependencies into feature extraction, and improving prediction performance on diverse datasets.

As the field matured, researchers began to incorporate more sophisticated comparisons. The authors of BrST-Net[32] extended ST-Net by analyzing various feature extracting techniques, while SEPAL[33] evaluated different approaches for predicting gene expression: global for the whole image, patch-based local, and spatially-informed, introducing the need of context in the prediction. These advances led to the development of other graph-based methods that considered the local context, such as THItoGene[34], EGN[35], or ErwaNet[36].

Advances in the AI imaging community enabled leveraging the multi-resolution nature of H&E images. M2ORT[37] and TRIPLEX[38] proposed multi-scale feature extractors, achieving notable correlation improvements across various datasets. And, recently, the authors of BLEEP[39] explicitly trained a bimodal embedding-based framework for spatial transcriptomics prediction from H&E.

Learning features correlated with gene expression

The general task for morphology translation methods is to input an image patch and output the gene expression patterns in that area. This task requires a series of design choices to effectively capture and predict the complex relationships between morphological features and gene expression.

Selection of training genes. Selecting which genes the model should learn is crucial because sequencing-based spatial transcriptomics methods typically cover the entire transcriptome, encompassing nearly 20,000 genes. However, many of these genes do not exhibit spatial patterns or have low expression levels. To ensure the learned features are meaningful and fall within relevant quadrants of our framework, careful gene selection is necessary.

Three main approaches, set by the first works, have been established for gene selection, which have the same goal are as follows:

Selecting the genes with the highest mean expression across the dataset, with the rationale that these genes are likely to be more robust and exhibit clearer spatial patterns. ST-Net and BrST-Net utilized this approach.

BOX 1

Formal framework definition

Let $\mathcal{D} = \{(x_G^1, x_M^1, y^1), \ldots, (x_G^N, x_M^N, y^N)\}$ be the paired multi-modal dataset of two modalities (gene expression G and morphology M) and N number of data points with spatial coordinates $\{(i^1, j^1), \ldots, (i^N, j^N)\}$. y^n represents the potential task-relevant information that is contained at (i^n, j^n).

The spatial transcriptomics workflow usually consists of processing the data until we obtain, per spatial coordinate, a feature corresponding to the point- or aggregate-gene expression in the local area. For the sake of notation, we represent these features as the result of encoding the spatial transcriptomics data with a gene expression-specific encoder $h_G = e_{\theta_G}(x_G)$. In parallel, we can represent the morphological features obtained from the paired images by an image-specific encoder $h_M = e_{\theta_M}(x_M)$.

Finally, based on the ideas of minimal sufficient statistics[96] and information bottleneck theory[97,98] proposed by Tian et al.[99], we define $I(\boldsymbol{a}; \boldsymbol{b})$ as the shared information between $\boldsymbol{a}$ and $\boldsymbol{b}$.

We can assume that the amount of task-relevant information contained in the gene expression is fixed and not negligible:

$$I(h_G; y) > 0 \tag{9}$$

The only thing we can control is how we train our feature extractor for the morphology data. Thus, we assume that we can only control e_{θ_M}, the morphology-specific encoder, and the only relationship we can capture is $I(\boldsymbol{h_G}; \boldsymbol{h_M})$.

Undoubtedly, we always want the morphological features to be as informative as possible:

$$\max_{\theta_M} I(h_M; y) \tag{10}$$

But depending on how the morphological feature descriptor is trained, we can end up with four different scenarios. We interpret these four scenarios by the four quadrants, as depicted in Fig. 1b.

Translation. We want to ensure that the morphology descriptors contain the maximum task-relevant information. For this, we can use the information included in gene expression as a proxy for relevance. This would require to maximize the shared information between both modalities while maintaining the relevance of morphology:

$$\max_{\theta_M} I(h_G; h_M) \tag{11}$$

so that the information shared between modalities is relevant, achieving the sweet point[99]:

$$I(h_G; h_M) = I(h_G; y) = I(h_M; y) \tag{12}$$

Integration. We want to ensure that the fusion, by a fusion module f_ψ, between the modalities carries more information than the individual modalities, minimizing redundancies between modalities, without adding nuisance information.

$$\min_{\theta_M} I(h_G; h_M) \tag{13}$$

so that the addition of morphology to gene expression carries more task-relevant information than gene expression alone:

$$I(f_\psi(h_G, h_M); y) > I(h_G; y) \tag{14}$$

Selecting the most highly variable genes across different spatial locations, with the rationale that these genes are more likely to capture meaningful spatial differences. Methods like HisToGene, Hist2ST, THItoGene, M2ORT, TRIPLEX, EGC, and ErwaNet have adopted this approach, aiming to enhance the models' ability to learn diverse and informative patterns.

Finally, manual selection of genes based on prior biological knowledge or specific research goals allows for the inclusion of genes known to be relevant to particular conditions or tissues, ensuring that the selected genes are biologically significant. DeepSpaCE, STimage, and BLEEP used manual selection to tailor their models to specific biomarkers of interest.

These approaches highlight that current methods are not yet capable of predicting the entire transcriptome due to the excessive noise present in such highly dimensional data.

Training regimes. Once we have selected our training genes, the next crucial step is to choose how to learn morphological features that correlate with these genes. The general approach involves a straightforward deep learning regression pipeline: input–model–output. Here, the input consists of image patches, and the output is the gene expression patterns for those patches. The main variable in this setup is the type of model used.

Early methods in this field leaned heavily on CNNs due to their proven effectiveness in image analysis. These models excel in capturing spatial hierarchies through their convolutional layers. For instance, ST-Net employed DenseNet-121[40]. DeepSpaCE, another pioneering method, utilized the simpler VGG16[41], which offered a straightforward yet powerful tool for feature extraction. BrST-Net expanded on this approach by comparing various CNN architectures and ultimately identified EfficientNet[42] as the top performer. Similarly, STimage utilized ResNet50[43] as the backbone for feature extraction but incorporated negative binomial regression layers to model gene expression as a distribution.

One significant shift was towards Transformer-based models, which are adept at capturing long-range dependencies in data. HisToGene was an early adopter of the Vision Transformer (ViT)[44], which applies the transformer architecture directly to image patches, treating them as sequences of tokens. This ability to capture global context more effectively than traditional CNNs was

https://doi.org/10.1038/s41467-025-58989-8

BOX 2

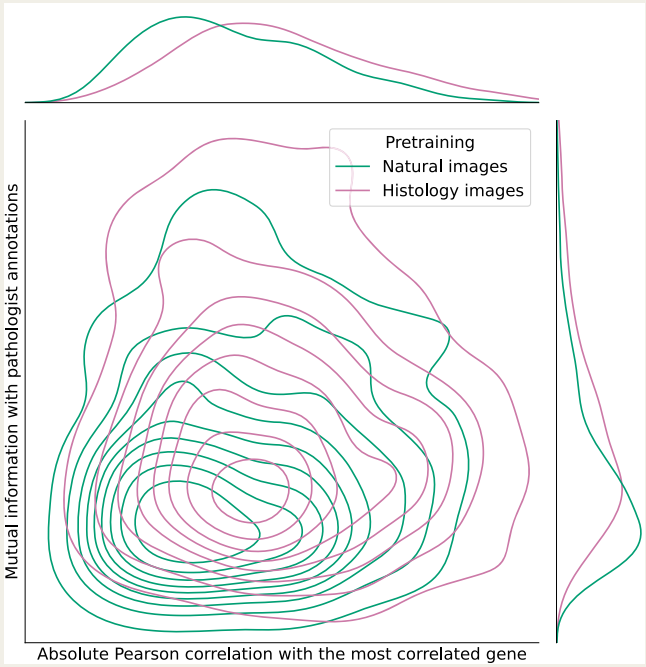
Practical example for the framework

We aim to visualize these four quadrants in practice by using proxies to measure the relevance of morphological features and their shared information with gene expression.

Relevance, in this context, is assessed using pathologist annotations on H&E slides as the ground truth. By applying a mutual information-based feature selection approach, we quantify the degree to which each morphological feature aligns with these annotations. A higher placement on the *y*-axis indicates greater relevance to the pathologist's annotations, reflecting the feature's task-specific importance. For *shared information* with gene expression, we propose using the absolute Pearson correlation between each morphological feature and its most correlated gene expression. This measures how well a morphological feature captures spatial patterns that align with specific gene expressions.

To illustrate these concepts, we compare two feature extraction scenarios using a ResNet-18 model pretrained in two different ways: on ImageNet[68] and via self-supervised learning on millions of H&E slides[100]. Both models extract 512 features from the penultimate layer. The plot below shows a shift in the feature distribution when using self-supervised pretraining on histology images, moving features further from the third quadrant (noise) and closer to the first quadrant (translation). This suggests that self-supervised pretraining on histology images captures more task-relevant features that align better with gene expression patterns.

Combining multiple features could improve predictive power. However, this analysis serves as a simplified demonstration of how the framework can be applied to real data using straightforward proxy metrics. The intention is not to claim that a single feature is sufficient, but to illustrate the general methodology for quantifying feature relevance and correlation. A more comprehensive analysis, combining multiple features (e.g., via PCA or other dimensionality reduction techniques), would likely enhance the predictive power, but for simplicity and interpretability, we focused on the highest correlation values in this example.



beneficial for understanding complex spatial patterns in gene expression data.

To further enhance the understanding of spatial organization, researchers introduced GNNs[45]. SEPAL combined the strengths of ViTs for obtaining image embeddings with GNNs to learn spatial patterns. This approach leveraged the global feature capturing of transformers and the spatial relationship modeling of GNNs, providing a comprehensive framework for morphology translation. Similarly, THItoGene used Efficient-Capsule Networks[46], designed to better capture spatial hierarchies, to generate embeddings, and combined ViTs and GNNs for reconstructing gene expression. Hist2ST also combines Convmixers for local feature extraction, transformer modules for capturing global spatial dependencies, and GNNs to model local spatial relationships.

The pyramidal nature of histopathological images, with features at multiple scales, led to the development of multi-level models. M2ORT, for example, used a hierarchical ViT[47] to accommodate the multi-scale nature of these images. By processing images at different scales, the hierarchical ViT could capture fine-grained details as well as broader contextual information, making it particularly effective for gene expression prediction. TRIPLEX adopted a similar approach with a multi-resolution architecture based on ResNet-18[43].

Recent advancements introduced bimodal embedding-based frameworks, which aim to create joint representations of images and gene expression data. BLEEP was the first to explicitly train such a framework, similar to CLIP[48].

Training and test splits. The choice of training and test splits is critical for the validation of the generalization abilities of deep learning models, especially given the typically small size of these datasets. Optimally, the methods should be evaluated on a completely independent test set never seen during training.

https://doi.org/10.1038/s41467-025-58989-8

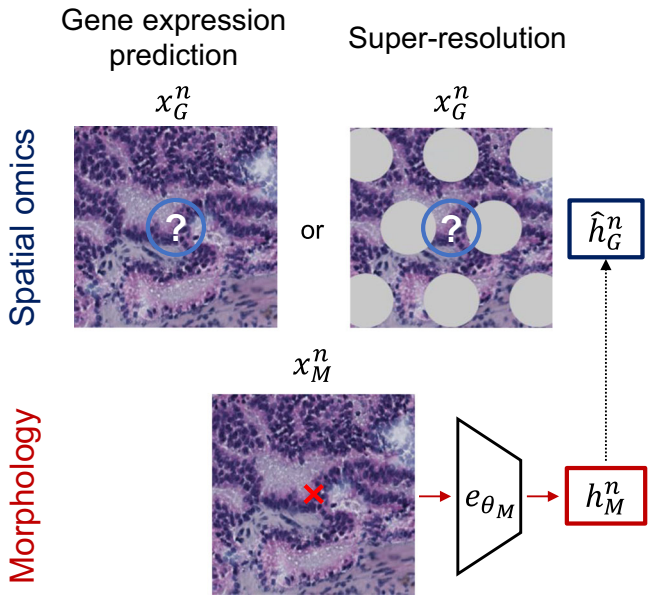


Fig. 2 | Morphology translation tasks. The most common task involving morphology translation is the prediction of gene expression. This can be done either by inferring the gene expression in a new sample or by imputing the gene expression in between areas containing gene expression.

However, a common approach is leave-one-out cross-validation. It involves using each sample as the test set, one at a time, while the remainder is used as a training set. While it maximizes the use of the dataset, it provides a limited view of the generalization of the model. ST-Net, HisToGene, Hist2ST, and THitoGene have employed leave-one-out to assess their models.

Another approach for small datasets is *k*-fold cross-validation, where the dataset is divided into *k* subsets and, one at a time, is used as a test and the remaining for training. This ensures that the performance is at least averaged across different partitions, and it is the approach of BrST-Net, TRIPLEX, EGN, and ErwaNet.

Finally, methods such as SEPAL, BLEEP, and M2ORT use traditional train–validation–test partitions, which can suffer from bad splitting choices but show generalization ability on independent samples. STimage, BrST-Net, and TRIPLEX validated their method using leave-one-out and k-fold cross-validation, but additionally tested on an independent sample.

Other tasks involving morphology translation

Although super-resolution can be obtained by dense prediction of gene expression, some authors have proposed other methods to achieve this. The earliest work published actually achieving pixel-level super-resolution spatial transcriptomics from H&E is, to our knowledge, XFuse[49]. At the time, super-resolution lacked a ground-truth benchmark for comparison, prompting the authors to demonstrate the robustness and potential of their method by comparing their results with in situ hybridization. iStar by Zhang et al.[50] is another super-resolution method that aims at predicting gene expression at the super-resolution level. The newer Xenium technology[51] with higher gene expression resolution enabled a quantitative comparison with XFuse, showing that their method enabled predictions that were closer to the ground truth.

One of the few instances where translation by deep learning was made from DAPI images instead of H&E was demonstrated by us[52]. We utilized imaging-based spatial transcriptomics to annotate marker genes, training CNNs to classify tissue morphology. This adaptation indicates the versatility of deep learning methods in handling different imaging types for morphology translation. DAPI imaging data is often neglected but holds a lot of potential, as shown in ref. 53. BIDCell[54] also makes use of DAPI morphology and imaging-based spatial transcriptomics for self-supervised segmentation of subcellular spatial transcriptomics data.

We also explored the correlation of latent features in different networks with gene expression[55]. We discovered that networks trained for cancer classification tasks[56] already contained features correlating with genes associated with prostate cancer[57]. Building on this, we introduced MHAST, a framework that employs self-supervised features to guide the re-assignment of deconvolved spatial transcriptomics[58]. Our validation on Tangram[59] demonstrated that MHAST could retrieve more accurate cell instances compared to original random allocations, showcasing its potential for achieving cell-level resolution.

More recently, Gao et al. presented IGI-DL[60], a method that uses predicted gene expression from H&E to predict patient prognosis. This is done by generating a graph, with predicted gene expression as nodes, to develop a survival model that outperforms others in breast and colorectal cancer cohorts. ELD[61] can also be considered a translation method as it can find matching points between imaging and spatial transcriptomics to align them together, making it easier to combine their information.

Morphology integration for spatial domain identification

Morphology integration involves identifying morphological features that spatially complement gene expression patterns. The primary application in this scenario is spatial domain identification. The goal is to learn morphological features that are not correlated with gene expression but still add additional information to define meaningful regions (see Fig. 3).

To the best of our knowledge, SpaCell[62] is the first method that combines morphology with spatial gene expression. They propose a pretrained CNN to extract morphology that was then combined with gene expression to obtain a joint latent space. This integrated space was then used for clustering and domain identification, setting a foundational precedent for future studies.

Hu et al. proposed SpaGCN[63], a graph-based method to further integrate spatial location with gene expression and histology. In this case, they use RGB intensity values as morphology descriptors and, through iterative clustering of the resulting graph, they obtain spatial domains.

The still-unpublished conST[64] proposes a contrastive approach for integrating gene expression, spatial information, and morphology. The authors use a pretrained autoencoder to obtain histology representations, which they then combine with gene expression in a GNN to obtain common features. The authors also implemented an interpretability module to explore the correlation between spots.

In a similar fashion as SpaCell, stLearn[65] also uses a pretrained CNN as a feature extractor. Together with the spot location, these features are used for normalization to adjust the gene expression values. They use this adjusted gene expression for tasks such as trajectory inference of cell-cell interaction analysis.

More recently, ConGI[66] proposed contrastive learning for obtaining a joint representation of histology and gene expression. The authors use a pretrained network as an image encoder and three different contrastive losses to model the relationships between the two modalities.

MorphLink[67] is one of the few integration methods that uses features at the single-cell level. It utilizes classical image analysis techniques to segment the patches and try to obtain interpretable features. The method then calculates a similarity score between the gene

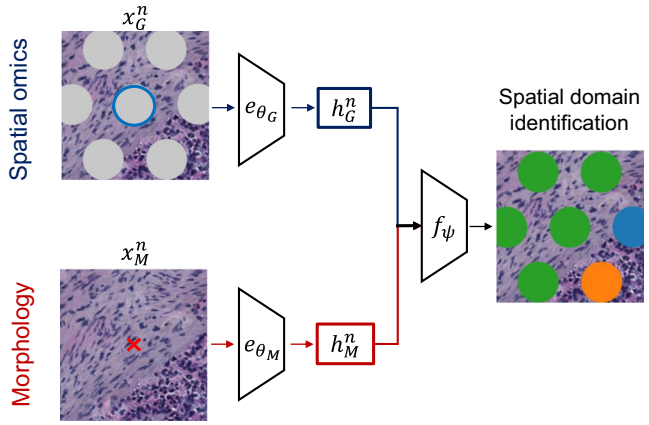


Fig. 3 | Morphology integration tasks. The most common task involving morphology integration is the identification of spatial domains, which can then be used in downstream tasks.

expression and the obtained features, much like our proposed workflow, but focuses on their shared patterns.

Extracting features that complement gene expression

A good morphological feature extractor for integration ensures that the obtained features have complementary information to gene expression and do not add noise, so as to obtain more meaningful domains compared to using gene expression alone.

Using CNN pretrained on large datasets like ImageNet[68] is by far the most common approach to obtain morphological descriptors for integration and other tasks. Authors typically use the penultimate layer before prediction. It is the case for SpaCell and stLearn, which use a pretrained ResNet50[43] and ConGI, which uses DenseNet-121[40]. conST employed an ImageNet pretrained masked autoencoder[69] to obtain morphological features. In contrast, SpaGCN and MorphLink do not use any deep learning approach to obtain features from the patches.

Except for MorphLink, which specifically maximizes the agreement between morphology and gene expression, these features are not explicitly trained to capture gene expression patterns, so one could argue that they probably fall in quadrants I and III of Fig. 1b. Authors probably expect that these features are general enough to add morphological information that is not already contained in the gene expression. But it is fair to argue that networks trained on natural images or RGB values might not provide features relevant enough for the integration task[53]. Recent advances in foundation models can promise to improve the feature extraction step, as we discuss in outlook and perspectives below.

Fusion modules for morphological and gene expression features

Apart from the choice of image encoder, it is important to decide how to integrate the modalities: the fusion module. It is worth noting that all integration methods use modality-specific encoders, due to the fixed workflow established for spatial transcriptomics, but a transformer-based architecture to handle input tokens for both modalities and use the same encoder for both is an interesting avenue[70].

No explicit fusion. SpaGCN does not have an explicit fusion module, instead, it constructs a graph with molecular features as the nodes and a combination of the spatial location and RGB image intensity features as the distance. stLearn also does not use a module for fusing the data and instead uses the 50 principal components (PCs) from the 2048 features pretrained ResNet50 together with the spatial locations to normalize the gene expression.

Representation stitching. SpaCell concatenates the latent representations of two auto-encoders that input the 2048 last layer features from the pretrained ResNet50 on the imaging side and the 2048 top highly variable genes from spatial transcriptomics.

Representation fusion. ConST generates a graph with KNN of spatial coordinates as edges and a concatenation of the 768 features extracted from the pretrained masked autoencoder and the 300 PCs from the genes as node features. It then generates the lower-dimensional graph with them by self-supervised learning. ConGI also concatenates the representations from a gene expression MLP encoder and a DenseNet-121 CNN for the images and feeds them into a small neural network to reduce the dimensionality in a self-supervised way as well.

Evaluation metrics, datasets, and benchmarks

Metrics for assessing the learned representations

Selecting the best possible representation of tissue morphology, both for translation and integration tasks, requires metrics to evaluate how well the learned features may solve the task at hand. The metrics introduced by the pioneering publications in the field are still widely used, despite their limitations[12,13]. Figure 4 shows an overview of these commonly used metrics.

Metrics for assessing morphological translation. In our framework, the ideal metric would measure how well the morphological features describe the gene patterns but, ideally, also include some information about their biological relevance. Authors have typically circumvented this by either selecting meaningful genes for prediction or analyzing the predictions ad hoc. However, there is currently no way of assessing this in a quantitative manner, so many studies achieving high-performance metrics might be in the overestimation quadrant of the framework, having good results on non-informative genes.

Pearson's correlation coefficient. Pearson's correlation coefficient (PCC), also represented by ρ, is a measure of the linear relationships between the observed and the predicted gene expression. Mathematically, it is computed as the covariance (cov) of the two variables divided by the product of their standard deviations (σ). The covariance is a measure of the tendency between the observed and predicted expression, i.e., if higher true expression levels match higher predicted levels, the covariance is higher. The division by their σ is a way of normalizing so that the metric is bound between +1 perfect positive and −1 perfect negative correlation.

Let y be the observed true gene expression for a gene, $\hat{y}$ be the predicted gene expression for that gene, and σ_y and $\sigma_{\hat{y}}$ their standard deviations, then the PCC can be expressed by:

$$\mathrm{PCC} = \frac{\mathrm{cov}\,(y, \hat{y})}{\sigma_y \sigma_{\hat{y}}} \tag{1}$$

Regression metrics. The problem of gene expression prediction is typically framed as a regression problem. More recent methods also report the performance of different regression metrics, namely mean absolute error (MAE), mean square error (MSE), and its root (RMSE). Let y_i be the observed true gene expression for a gene at position i, and $\hat{y}_i$ be the predicted gene expression for that gene at the same position. In other words, a small error means that we have learned features that predict gene expression well.

The MAE measures the average magnitude of the errors in a set of predictions, without considering their direction. It is the average over the test sample of the absolute differences between prediction and

https://doi.org/10.1038/s41467-025-58989-8

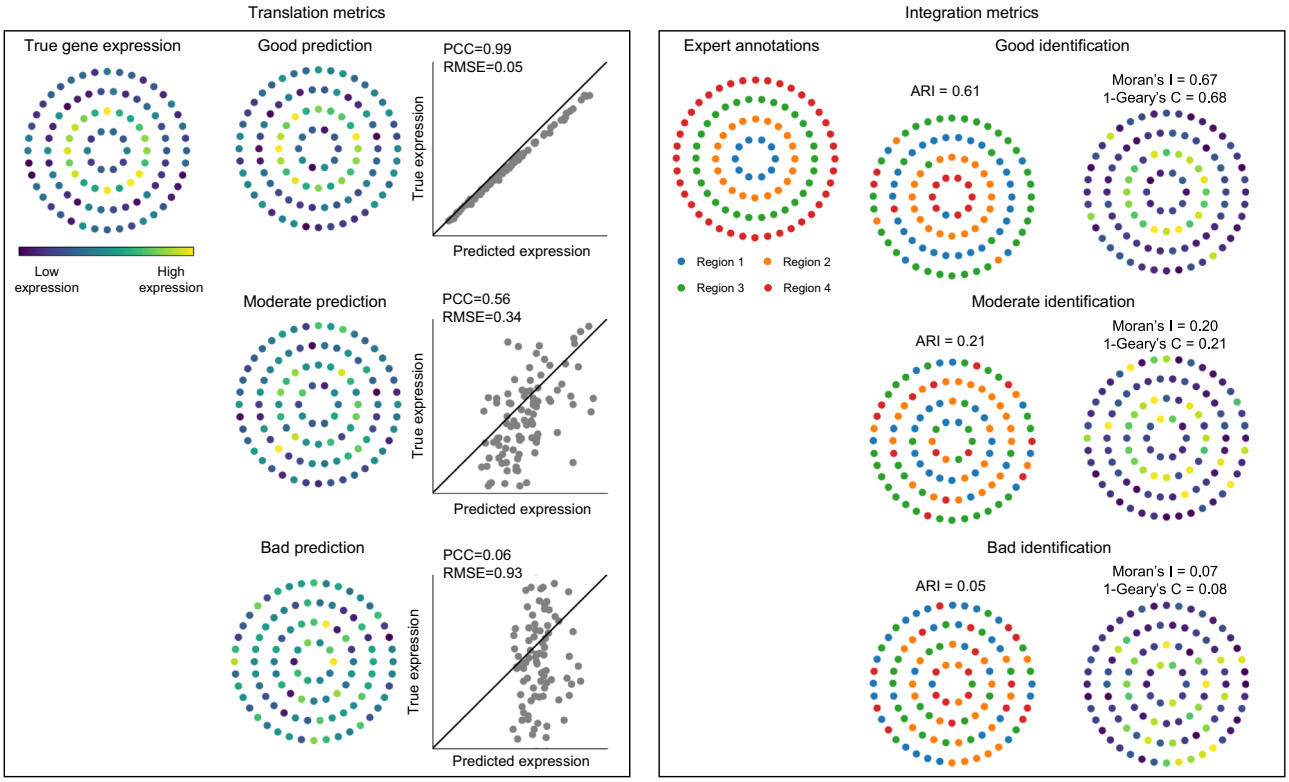


Fig. 4 | Commonly used metrics for morphology translation and integration. This synthetic example presents tissue regions as concentric circles. **Left:** Translation metrics usually measure the agreement of the true gene expression with the gene expression predicted from morphology. Pearson's correlation coefficient (PCC) and regression metrics such as the root mean squared error (RMSE) are usually employed. **Right:** Integration metrics measure the agreement of expert annotations with the domains defined jointly by morphology and spatial transcriptomics via the adjusted Rand index (ARI). It is common to also define spatially variable genes that define the identified domains and measure their degree of spatial autocorrelation with Moran's *I* or Geary's *C*.

actual observation, where all individual differences have equal weight.

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{2}$$

The MSE measures the average of the squares of the errors, that is, the average squared difference between the estimated values and the actual value. MSE is a risk function corresponding to the expected value of the squared error loss.

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{3}$$

The RMSE is the square root of the average of squared differences between the prediction and the actual observation. It provides a measure of how spread out these residuals are. In other words, it tells you how concentrated the data is around the line of best fit.

$$\mathrm{RMSE} = \sqrt{\mathrm{MSE}} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

Each of these metrics offers a different perspective on the accuracy of the model predictions. MAE gives an idea of the magnitude of errors, MSE emphasizes larger errors more than smaller ones due to squaring the differences, and RMSE provides an overall measure of fit in the same units as the response variable. However, the values are not bounded, which can be harder to interpret and compare.

Metrics for assessing morphological integration. In our framework, the ideal metric for identifying strong features for morphological integration would measure how the morphological features, despite not being correlated with gene expression patterns, are still relevant and thus add additional information to the spatial domain identification. On the one side, authors typically measure how well their clustered domains agree with expert annotations. On the other hand, the defined domains can be linked to a specific spatially variable gene (SVG). This gene can be validated by its amount of autocorrelation.

Clustering evaluation metrics. Since spatial domain identification is not a classification problem in which there is a one-to-one connection between regions, we need metrics that take this into account. The Rand index (RI) and its adjusted version (ARI)[71–73], measure the similarity of two clusters, accommodating for scenarios with unlabeled and different amounts of categories. To evaluate the similarity between two clustering results, we use metrics that account for such complexities.

The ARI is a measure of the similarity between two data clusterings, correcting for the chance grouping of elements. The ARI ranges from −1 to 1, where 1 indicates perfect agreement between the two clusterings, 0 indicates random clustering and negative values indicate less agreement than expected by chance.

$$\mathrm{ARI} = \frac{\mathrm{RI} - \mathrm{Expected\ RI}}{\max(\mathrm{RI}) - \mathrm{Expected\ RI}} \tag{5}$$

where RI is the measure of the similarity between two data clusterings, and Expected RI is the expected value of the RI for random clustering.

Autocorrelation metrics. Gene expressions at different locations may exhibit spatial autocorrelation, where nearby locations have similar expression levels. To assess this, authors use Moran's I[74,75] and Geary's C[76,77] statistics.

Moran's I quantifies the overall spatial autocorrelation of gene expression. It ranges from −1 to 1, where 1 indicates a clear spatial pattern, 0 indicates random spatial expression, and −1 indicates a chessboard-like pattern.

$$\text{Moran's } I = \frac{N}{W}\frac{\sum_i \sum_j w_{ij}(x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2} \tag{6}$$

where x_i and x_j are the gene expressions at spots i and j, $\bar{x}$ is the mean expression, N is the total number of spots, w_{ij} is the spatial weight between spots i and j, and W is the sum of w_{ij}. A common choice it to set $w_{ij} = 1$ for the 4 nearest neighbors of spot i and $w_{ij} = 0$ otherwise.

Geary's C also measures spatial autocorrelation but focuses on local differences. Its value ranges from 0 to 2.

$$\text{Geary's } C = \frac{(N-1)}{2W}\frac{\sum_i \sum_j w_{ij}(x_i - x_j)^2}{\sum_i (x_i - \bar{x})^2} \tag{7}$$

To align it with Moran's I, we scale it to [−1, 1]:

$$\text{Scaled Geary's } C = 1 - \text{Geary's } C \tag{8}$$

Here, 1 indicates perfect positive autocorrelation, 0 indicates no autocorrelation, and −1 indicates perfect negative autocorrelation. Both metrics provide insight into the spatial patterns of gene expression.

Public datasets and benchmarks

As with the metrics, the datasets that the community has used for developing new methods draw inspiration from the first works. The scarce availability of big spatial transcriptomics datasets further enforced this.

Many translation methods use breast cancer datasets, either the two ductal carcinoma samples from 10X Visium and the 23 breast cancer patients used by ST-Net[27] or the 32 HER2-positive breast cancer samples presented by Andersson et al.[52]. Another commonly used dataset has been the human squamous cell carcinoma dataset[78]. This is also the case for the benchmark presently available for translation studies[12]. Even though the authors also include methods that do not train on spatial data, they analyze the performance of ST-Net, DeepSpaCE, and HisToGene on the HER2-positive and squamous cell carcinoma datasets.

Super-resolution tasks are typically harder to evaluate quantitatively, as there is no ground truth for per-pixel gene expression. New technologies such as 10X Xenium were utilized as a surrogate for this by iStar and MHAST. This enabled the authors to have a denser gene expression map that could be then compared to their methods, even though 10X Xenium is an imaging-based method and thus requires predefining a gene panel of 100s of genes, far from the whole transcriptome.

The two main datasets used for integration methods are 10X Visium mouse brain samples and the 10X Visium human dorsolateral prefrontal cortex available at the spatialLIBD library[79]. The latter was used in the benchmark by Yuan et al.[13]. As the integration task is commonly spatial domain identification, it requires some type of annotations for comparison, limiting the amount of available datasets.

Technical considerations

Working with histological images comes with its own set of challenges that are far from solved[80,81]. One of the primary issues is the inherent variability in histological image data, which arises from differences in staining protocols, sample preparation techniques, and imaging conditions across different laboratories and studies. This variability complicates the generalization of models, as algorithms trained on one dataset may not perform well on another due to these inconsistencies. Pretraining models on a wide variety of data from different sources to generate so-called foundation models can help improve their robustness and generalization capabilities[82].

To this, we add the challenges of working with such high-dimensional data as spatial transcriptomics. Each modality individually produces complex, high-dimensional data, and integrating these datasets compounds the complexity. Dimensionality reduction techniques are usually employed to reduce the complexity of the data while trying to retain the essential information[83,84].

Another challenge in spatial transcriptomics is the alignment of morphology features with varying resolution levels, from multi-cellular to single-cell and subcellular scales, as seen in technologies like 10x Xenium[51]. These differences require methods to align morphology from the commonly used patches with spatial transcriptomics data at different resolutions. To address this, techniques such as rescaling or interpolation are used to ensure consistency across modalities. When integrating subcellular-resolution data with morphology, it's crucial to maintain alignment without introducing discrepancies, which can be achieved through multi-resolution fusion methods or hierarchical modeling strategies.

These challenges make it hard to benchmark and compare methods effectively. Many authors compare their methods against previous work, often reporting consistent improvements. However, these comparisons are frequently made using new datasets and varying validation schemes, which complicates meaningful comparisons. The limited availability of large, diverse datasets exacerbates this issue, as it hinders the ability to generalize results and often leads to contradictory findings when compared to the original authors' reports[12,13]. Since spatial transcriptomics data can be highly variable based on tissue source and organ, a broader collection of annotated data, such as[85] is essential for both method development and comparison. Transparent reporting of training and validation procedures, along with benchmarking across diverse real-world datasets, is crucial for ensuring reproducibility and establishing reliable comparisons across studies.

An important consideration for our study is the reassessment of the metrics currently utilized, with the possibility of incorporating an additional dimension. Traditional translation tasks often employ PCC or regression metrics, which primarily evaluate the shared information between learned morphological features and gene expression patterns, analogous to the x-axis in our proposed framework. This approach can lead to an overestimation of predictive performance for genes that are not task-relevant, such as housekeeping or constitutive genes, thus neglecting genes that may hold significant clinical value. Similarly, integration methods typically assess the shared information between the fusion of morphological features and gene expression with expert annotations, comparable to the y-axis in our framework. This methodology can entangle the individual contributions of each modality and potentially diminish performance by incorporating noisy or correlated morphological features.

Finally, an often overlooked aspect is the extensive training and computational resources required in these applications. In predicting gene expression, especially with internal validation as opposed to an independent test set, additional training can substantially enhance results. This concern also extends to integration tasks, where it becomes difficult to determine whether improved performance is attributable to the actual contribution of morphological data or merely the result of over-fitting to the specific problem.

Outlook and perspectives

The combination of morphology and spatial transcriptomics is at its dawn. In this work, we presented a mental framework to conceptualize

and guide future developments in this field, as many challenges remain unresolved.

Correlation values for gene expression prediction are rather moderate. Even with modern implementations, performances are far from clinically transferable. If we want to be able to apply such methods to predict the gene expression in large H&E-stained cohorts, we have to make sure we are falling in quadrant I of Fig. 1b. Achieving this will require not just an improvement in average model performance, but a focus on ensuring that methods are truly effective for clinically important genes. This might involve developing specialized models tailored to specific clinical problems, ensuring their applicability in real-world scenarios.

Spatial domain identification has not yet shown significant benefits from morphological integration. This is likely due to the fact that the morphological features currently used are either not relevant enough or provide redundant information already captured by gene expression data. Future method development should prioritize the creation of morphological descriptors that are both highly relevant and complementary to gene expression, ensuring that they add value rather than noise to the analysis. Until such improvements are made, the integration of morphology should be approached with caution, avoiding the inclusion of irrelevant or redundant features.

Foundation models for histopathology, trained on vast datasets using self-supervised learning, have demonstrated remarkable potential as feature extractors or fine-tuned tools for specific tasks. Rapid advancements in 2024 have led to the development of models like CONCH[86], CTrans-Path[87], H-optimus-0[88], Kaiko[89], Phikon[90], Prov-GigaPath[91], UNI[92], or Virchow[93], showcasing state-of-the-art representation learning. Despite these strides, their integration with spatial transcriptomics remains limited, as the field has yet to effectively leverage these representations to enhance spatial domain identification or gene expression prediction. This requires concerted efforts to align morphological insights from foundation models with spatial transcriptomics, prioritizing clinical relevance and complementarity to gene expression data.

Another interesting but poorly explored aspect of morphology integration is that of time. The Nilsson lab presents an in situ sequencing-based workflow that generates detailed quantitative maps of genetic subclone composition across whole-tumor sections, thereby visualizing a 'pseudo-timeline' of tumor progression, along with morphological changes[94]. A form of temporal data is also provided in ref. 95, where cell-type signatures are modeled over space and time in the developing human heart. A more in-depth investigation of the relationship between gene expression and morphology, combined with these proxies of time and space, is still to be done. Furthermore, tissue morphology may hold a memory of genes that were expressed early during the tissue development, but are no longer expressed. And, at the same time, gene expression is likely to precede changes in morphology, holding the power to predict changes likely to occur in the near future. Integration may thus, for example, help us identify very early signs of upcoming changes in morphology, and this discordance in time may be one explanation for limited correlation in translation efforts.

The framework proposed in this review extends beyond the integration of morphology and spatial transcriptomics and can be adapted for multi-modal or multi-omics integration tasks. In the case of translation, the goal remains to maximize shared information between modalities while focusing on the specific downstream task. For integration, the aim is to minimize redundancy while ensuring meaningful contributions from both modalities. As spatial multi-modal technologies, such as those combining gene expression and the epigenome, become more prevalent, the framework can be expanded to accommodate these complex scenarios. In such cases, the interaction between multiple omics spaces and the morphology space becomes crucial, with each omics modality contributing a fixed set of relevant information. The morphology space, however, remains flexible, moving across different axes of interaction. This extension of the framework provides a more robust approach for integrating multiple omics modalities and could lead to the development of new translation scenarios that reflect the intersection of various omics data with morphology.

Ultimately, we hope this framework will serve as a compass for researchers, facilitating the joint use of morphology and spatial transcriptomics as imaging AI and bioinformatics continue to advance. By providing a structured approach to integrating these complex data types, this framework aims to guide efforts to gain deeper insights into biological processes and disease mechanisms.

References

1. Longo, S. K., Guo, M. G., Ji, A. L. & Khavari, P. A. Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics. *Nat. Rev. Genet.* **22**, 627–644 (2021).
2. Rao, A., Barkley, D., França, G. S. & Yanai, I. Exploring tissue architecture using spatial transcriptomics. *Nature* **596**, 211–220 (2021).
3. Bressan, D., Battistoni, G. & Hannon, G. J. The dawn of spatial omics. *Science* **381**, eabq4964 (2023).
4. Park, J. et al. Spatial omics technologies at multimodal and single cell/subcellular level. *Genome Biol.* **23**, 256 (2022).
5. Seferbekova, Z., Lomakin, A., Yates, L. R. & Gerstung, M. Spatial biology of cancer evolution. *Nat. Rev. Genet.* **24**, 295–313 (2023).
6. Zhao, T. et al. Spatial genomics enables multi-modal study of clonal heterogeneity in tissues. *Nature* **601**, 85–91 (2022).
7. Zhou, R., Yang, G., Zhang, Y. & Wang, Y. Spatial transcriptomics in development and disease. *Mol. Biomed.* **4**, 32 (2023).
8. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
9. Cui, M. & Zhang, D. Y. Artificial intelligence and computational pathology. *Lab. Investig.* **101**, 412–422 (2021).
10. Song, A. H. et al. Artificial intelligence for digital and computational pathology. *Nat. Rev. Bioeng.* **1**, 930–949 (2023).
11. Lu, S., Fürth, D. & Gillis, J. Integrative analysis methods for spatial transcriptomics. *Nat. Methods* **18**, 1282–1283 (2021).
12. Wang, C. et al. Benchmarking the translational potential of spatial gene expression prediction from histology. *Nat. Commun.* **16**, 1544 (2025).
13. Yuan, Z. et al. Benchmarking spatial clustering methods with spatially resolved transcriptomics data. *Nat. Methods* **21**, 712–722 (2024).
14. Li, Y., Stanojevic, S. & Garmire, L. X. Emerging artificial intelligence applications in spatial transcriptomics analysis. *Comput. Struct. Biotechnol. J.* **20**, 2895–2908 (2022).
15. Zahedi, R. et al. Deep learning in spatially resolved transcriptfomics: a comprehensive technical view. *Brief. Bioinform.* **25**, bbae082 (2024).
16. Rodriques, S. G. et al. Slide-seq: a scalable technology for measuring genome-wide expression at high spatial resolution. *Science* **363**, 1463–1467 (2019).
17. Chen, A. et al. Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays. *Cell* **185**, 1777–1792 (2022).
18. Chen, K. H., Boettiger, A. N., Moffitt, J. R., Wang, S. & Zhuang, X. Spatially resolved, highly multiplexed rna profiling in single cells. *Science* **348**, aaa6090 (2015).
19. Eng, C.-H. L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH+. *Nature* **568**, 235–239 (2019).
20. Ke, R. et al. In situ sequencing for RNA analysis in preserved tissue and cells. *Nat. Methods* **10**, 857–860 (2013).

21. Larsson, C., Grundberg, I., Söderberg, O. & Nilsson, M. In situ detection and genotyping of individual mRNA molecules. *Nat. Methods* **7**, 395–397 (2010).
22. Janowczyk, A. & Madabhushi, A. Deep learning for digital pathology image analysis: a comprehensive tutorial with selected use cases. *J. Pathol. Inform.* **7**, 29 (2016).
23. Mondol, R. K. et al. hist2RNA: an efficient deep learning architecture to predict gene expression from breast cancer histopathology images. *Cancers* **15**, 2569 (2023).
24. Comiter, C. et al. Inference of single cell profiles from histology stains with the single-cell omics from histology analysis framework (SCHAF). *BioRxiv* 2023–03 (2023).
25. Wang, Y. et al. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression–morphology analysis in breast cancer. *Cancer Res.* **81**, 5115–5126 (2021).
26. Weitz, P. et al. Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression-based convolutional neural networks. *Bioinformatics* **38**, 3462–3469 (2022).
27. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat. Biomed. Eng.* **4**, 827–834 (2020).
28. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *BioRxiv* 2021–11 (2021).
29. Monjo, T., Koido, M., Nagasawa, S., Suzuki, Y. & Kamatani, Y. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Sci. Rep.* **12**, 4133 (2022).
30. Tan, X. et al. Stimage: robust, confident and interpretable models for predicting gene markers from cancer histopathological images. *bioRxiv* 2023–05 (2023).
31. Zeng, Y. et al. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief. Bioinform.* **23**, bbac297 (2022).
32. Rahaman, M. M., Millar, E. K. & Meijering, E. Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. *Sci. Rep.* **13**, 13604 (2023).
33. Mejia, G., Cárdenas, P., Ruiz, D., Castillo, A. & Arbeláez, P. SEPAL: spatial gene expression prediction from local graphs. In *Proc. IEEE/CVF International Conference on Computer Vision* 2294–2303 (IEEE 2023).
34. 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 (2024).
35. Yang, Y., Hossain, M. Z., Stone, E. & Rahman, S. Spatial transcriptomics analysis of gene expression prediction using exemplar guided graph neural network. *Pattern Recognit.* **145**, 109966 (2024).
36. Chen, C., Zhang, Z., Tang, P., Liu, X. & Huang, B. Edge-relational window-attentional graph neural network for gene expression prediction in spatial transcriptomics analysis. *Comput. Biol. Med.* **174**, 108449 (2024).
37. Wang, H. et al. M2ORT: many-to-one regression transformer for spatial transcriptomics prediction from histopathology images. *arXiv preprint arXiv:2401.10608* (2024).
38. Chung, Y., Ha, J. H., Im, K. C. & Lee, J. S. Accurate spatial gene expression prediction by integrating multi-resolution features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11591–11600 (IEEE, 2024).
39. Xie, R. et al. Spatially resolved gene expression prediction from histology images via bi-modal contrastive learning. *Adv. Neural Inf. Process. Syst.* **36** https://proceedings.neurips.cc/paper_files/paper/2023/hash/df656d6ed77b565e8dcdfbf568aead0a-Abstract-Conference.html (2024).
40. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (IEEE, 2017).
41. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In proceedings of International Conference on Learning Representations (ICLR), (2015).
42. Tan, M. & Le, Q. EfficientNet: rethinking model scaling for convolutional neural networks. In *Proc. International Conference on Machine Learning* 6105–6114 (PMLR, 2019).
43. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition* 770–778 (IEEE, 2016).
44. Dosovitskiy, A. et al. An image is worth 16x16 words: transformers for image recognition at scale. In *proceedings of International Conference on Learning Representations (ICLR)*, (2021).
45. Wu, Z. et al. A comprehensive survey on graph neural networks. *IEEE Trans. Neural Netw. Learn. Syst.* **32**, 4–24 (2020).
46. Mazzia, V., Salvetti, F. & Chiaberge, M. Efficient-CapsNet: capsule network with self-attention routing. *Sci. Rep.* **11**, 14634 (2021).
47. Chen, R. J. et al. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16144–16155 (IEEE, 2022).
48. Radford, A. et al. Learning transferable visual models from natural language supervision. In *Proc. International Conference on Machine Learning* 8748–8763 (PMLR, 2021).
49. Bergenstråhle, L. et al. Super-resolved spatial transcriptomics by deep data fusion. *Nat. Biotechnol.* **40**, 476–479 (2022).
50. Zhang, D. et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat. Biotechnol.* **42**, 1372–1377 (2024).
51. Janesick, A. et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nat. Commun.* **14**, 8353 (2023).
52. Andersson, A., Partel, G., Solorzano, L. & Wählby, C. Transcriptome-supervised classification of tissue morphology using deep learning. In *Proc. 2020 IEEE 17th International Symposium on Biomedical Imaging* 1630–1633 (IEEE, 2020).
53. Chelebian, E., Avenel, C. & Wahlby, C. Self-supervised learning for genetically relevant domain identification in morphological images. In *Proc. 2024 IEEE 21st International Symposium on Biomedical Imaging* (IEEE, 2024).
54. Fu, X. et al. BIDCell: biologically-informed self-supervised learning for segmentation of subcellular spatial transcriptomics data. *Nat. Commun.* **15**, 509 (2024).
55. Chelebian, E. et al. Morphological features extracted by AI associated with spatial transcriptomics in prostate cancer. *Cancers* **13**, 4837 (2021).
56. Ström, P. et al. Artificial intelligence for diagnosis and grading of prostate cancer in biopsies: a population-based, diagnostic study. *Lancet Oncol.* **21**, 222–232 (2020).
57. Erickson, A. et al. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature* **608**, 360–367 (2022).
58. Chelebian, E., Avenel, C., Leon, J., Hon, C.-C. & Wählby, C. Learned morphological features guide cell type assignment of deconvolved spatial transcriptomics. In *Proc. Medical Imaging with Deep Learning* (Journal of Machine Learning Research, 2024).

59. Biancalani, T. et al. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat. Methods* **18**, 1352–1362 (2021).
60. Gao, R. et al. Harnessing TME depicted by histological images to improve cancer prognosis through a deep learning system. *Cell Rep. Med.* **5**, 101536 (2024).
61. Ekvall, M. et al. Spatial landmark detection and tissue registration with deep learning. *Nat. Methods* **21**, 673–679 (2024).
62. Tan, X., Su, A., Tran, M. & Nguyen, Q. Spacell: integrating tissue morphology and spatial gene expression to predict disease cells. *Bioinformatics* **36**, 2293–2294 (2020).
63. Hu, J. et al. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods* **18**, 1342–1351 (2021).
64. Zong, Y. et al. conST: an interpretable multi-modal contrastive learning framework for spatial transcriptomics. *BioRxiv* 2022–01 (2022).
65. Pham, D. et al. Robust mapping of spatiotemporal trajectories and cell–cell interactions in healthy and diseased tissues. *Nat. Commun.* **14**, 7739 (2023).
66. Zeng, Y. et al. Identifying spatial domain by adapting transcriptomics with histology through contrastive learning. *Brief. Bioinform.* **24**, bbad048 (2023).
67. Huang, J. et al. MorphLink: bridging cell morphological behaviors and molecular dynamics in multi-modal spatial omics. *bioRxiv* 2024–08 (2024).
68. Deng, J. et al. ImageNet: a large-scale hierarchical image database. In *Proc. 2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (IEEE, 2009).
69. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition* 16000–16009 (IEEE, 2022).
70. Zong, Y., Mac Aodha, O. & Hospedales, T. Self-Supervised Multimodal Learning: A Survey. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, https://doi.org/10.1109/TPAMI.2024.3429301 (2023).
71. Rand, W. M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
72. Hubert, L. & Arabie, P. Comparing partitions. *J. Classif.* **2**, 193–218 (1985).
73. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: Is a correction for chance necessary? In *Proc. 26th Annual International Conference on Machine Learning* 1073–1080 (Association for Computing Machinery, New York, NY, United States, 2009).
74. Moran, P. A. Notes on continuous stochastic phenomena. *Biometrika* **37**, 17–23 (1950).
75. Li, H., Calder, C. A. & Cressie, N. Beyond Moran's I: testing for spatial dependence based on the spatial autoregressive model. *Geogr. Anal.* **39**, 357–375 (2007).
76. Geary, R. C. The contiguity ratio and statistical mapping. *Inc. Stat.* **5**, 115–146 (1954).
77. Anselin, L. A local indicator of multivariate spatial association: extending Geary's C. *Geogr. Anal.* **51**, 133–150 (2019).
78. Ji, A. L. et al. Multimodal analysis of composition and spatial architecture in human squamous cell carcinoma. *Cell* **182**, 497–514 (2020).
79. Pardo, B. et al. spatialLIBD: an R/bioconductor package to visualize spatially-resolved transcriptomics data. *BMC Genom.* **23**, 434 (2022).
80. Bera, K., Schalper, K. A., Rimm, D. L., Velcheti, V. & Madabhushi, A. Artificial intelligence in digital pathology-new tools for diagnosis and precision oncology. *Nat. Rev. Clin. Oncol.* **16**, 703–715 (2019).
81. Van der Laak, J., Litjens, G. & Ciompi, F. Deep learning in histopathology: the path to the clinic. *Nat. Med.* **27**, 775–784 (2021).
82. Moor, M. et al. Foundation models for generalist medical artificial intelligence. *Nature* **616**, 259–265 (2023).
83. Shang, L. & Zhou, X. Spatially aware dimension reduction for spatial transcriptomics. *Nat. Commun.* **13**, 7203 (2022).
84. Sun, Y. et al. A comprehensive survey of dimensionality reduction and clustering methods for single-cell and spatial transcriptomics data. *Brief. Funct. Genom.* **23**, 733–744 (2024).
85. Chen, J. et al. STimage-1K4M: a histopathology image-gene expression dataset for spatial transcriptomics. *arXiv preprint arXiv:2406.06393* (2024).
86. Lu, M. Y. et al. A visual-language foundation model for computational pathology. *Nat. Med.* **30**, 863–874 (2024).
87. Wang, X. et al. Transformer-based unsupervised contrastive learning for histopathological image classification. *Med. Image Anal.* **81**, 102559 (2022).
88. Saillard, C. et al. H-optimus-0. https://github.com/bioptimus/releases/tree/main/models/h-optimus/v0 (2024).
89. Aben, N. et al. Towards large-scale training of pathology foundation models. *arXiv preprint arXiv:2404.15217* (2024).
90. Filiot, A., Jacob, P., Mac Kain, A. & Saillard, C. Phikon-v2, a large and public feature extractor for biomarker prediction. *arXiv preprint arXiv:2409.09173* (2024).
91. Xu, H. et al. A whole-slide foundation model for digital pathology from real-world data. *Nature* **630**, 181–188 (2024).
92. Chen, R. J. et al. Towards a general-purpose foundation model for computational pathology. *Nat. Med.* **30**, 850–862 (2024).
93. Vorontsov, E. et al. A foundation model for clinical-grade computational pathology and rare cancers detection. *Nat. Med.* **30**, 2924–2935 (2024).
94. Lomakin, A. et al. Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature* **611**, 594–602 (2022).
95. Marco Salas, S. et al. De novo spatiotemporal modelling of cell-type signatures in the developmental human heart using graph convolutional neural networks. *PLoS Comput. Biol.* **18**, e1010366 (2022).
96. Soatto, S. & Chiuso, A. Visual representations: defining properties and deep approximations. *arXiv preprint arXiv:1411.7676* (2014).
97. Alemi, A. A., Fischer, I., Dillon, J. V. & Murphy, K. Deep variational information bottleneck. In *proceedings of the International Conference on Learning Representations (ICLR)*, (2017).
98. Tishby, N., Pereira, F. C. & Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057* (2000).
99. Tian, Y. et al. What makes for good views for contrastive learning? *Adv. Neural Inf. Process. Syst.* **33**, 6827–6839 (2020).
100. Ciga, O., Xu, T. & Martel, A. L. Self supervised contrastive learning for digital histopathology. *Mach. Learn. Appl.* **7**, 100198 (2022).

Acknowledgements

We would like to thank current and former collaborators and co-workers in the field of spatial transcriptomics and imaging AI for sharing their insights leading up to this review.

Author contributions

E.C.: conceptualization, formal analysis, investigation, methodology, visualization, and writing—original draft preparation. C.A.: conceptualization, resources, and writing—review and editing. C.W.: conceptualization, funding acquisition, supervision, and writing—review and editing.

Competing interests
C.W. is on the advisory board of Navinci Diagnostics, Sweden. The rest of the authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to Carolina Wählby.

Peer review information *Nature Communications* thanks Fan Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.