ActiveVisium: Leveraging Active Learning to Enhance Manual Pathologist Annotation in 10x Visium Spatial Transcriptomics Experiments

Jelica Vasiljević[1] (✉), Ines Berenguer Veiga[1], Kerstin Hahn[1], Petra Schwalie[1], and Alberto Valdeolivas[1]

Roche Pharma Research and Early Development, Roche Innovation Center Basel, Basel, Switzerland
- {jelica.vasiljevic, ines.berenguer_veiga, kerstin.hahn, petra.schwalie, alberto.valdeolivas}@roche.com

Abstract. Spatial transcriptomics (ST) technologies offer valuable insights into tissue organisation by capturing gene expression within its spatial context. Among these, 10x Visium stands out for its capacity to integrate gene expression profiles with histological images, facilitating multi-modal tissue analysis. However, comprehensive analysis requires manual pathologist's annotations at the capture spot level, a labour-intensive and time-consuming process that demands a significant amount of pathologists' time. Given the scale of studies involving multiple ST samples, manual annotation becomes impractical, and no automated solutions currently exist. To address this, we introduce ActiveVisium, an active learning framework designed to enhance spot-level annotation in 10x Visium datasets. To the best of our knowledge, ActiveVisium is the first framework to leverage tissue morphology and, optionally, gene expression data to automate large-scale spot annotation while selecting the most informative ones for manual labelling. Furthermore, this approach enables transfer learning across similar samples, thereby reducing annotation time for entire studies. Evaluations across breast cancer, colorectal cancer, and healthy kidney samples demonstrate that ActiveVisium has the potential to significantly improve annotation efficiency and consistency. All code and data are publicly available.

Keywords: : Spatial transcriptomics, Active learning, Digital Pathology, Deep Learning, 10x Visium

1 Introduction

Recent advancements in high-throughput technologies and imaging methods have enabled the development of ST, allowing for the capture of gene expression profiles within their native tissue context and opening new opportunities for investigating tissue organisation and function [17]. Various ST technologies are available [6], where some, such as 10x Visium [21], integrate gene expression

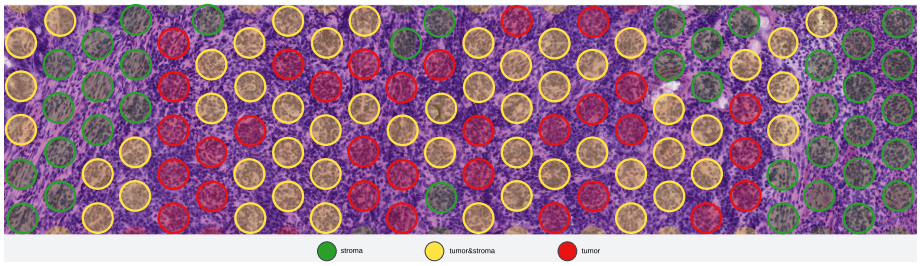

Fig. 1: Spot-level annotations in a FF colorectal cancer sample. Each spot, 55 μm in diameter, is labelled based on the tissue composition it covers. Standard 10x Visium slides have 5000 barcoded spots.

data with histological images, facilitating simultaneous analysis of molecular and morphological features of a tissue sample. With both modalities available—gene expression data and histological images—molecular analyses are often cross-referenced with pathologist annotations on the corresponding whole slide images (WSIs) [24,1]. Such a multi-perspective view of tissue is especially important for gaining a comprehensive understanding of tissue organisation, as some differences may only be visible at the molecular level. For example, in a Colorectal Cancer (CRC) study [24], spots with similar morphological features that were uniformly annotated as tumours by a pathologist exhibited distinct gene expression profiles.

The 10X Visium [21] platform is one of the most widely utilised ST technologies [30], compatible with both fresh-frozen (FF) and formalin-fixed, paraffin-embedded (FFPE) tissue sections. It allows for the capture of near-whole transcriptome readouts in specially designed barcoded spots, which can be mapped to a histological image of the tissue, as illustrated in Figure 1. The standard Visium platform employs capture slides containing approximately 5,000 spatially barcoded spots, each with a diameter of 55 μm. Additionally, slides with 11,000 spots are available. For a comprehensive analysis, annotations by pathologists should ideally correspond to individual capture spots, as illustrated in Figure 1.

To obtain relevant biological information, a typical experimental study design requires multiple samples under different conditions. This usually involves biological and, occasionally, technical replicates to draw statistically significant conclusions. Furthermore, each sample is processed in a separate Visium capture area, each requiring its own annotations for comprehensive analysis. In these conditions, the annotation task becomes highly repetitive, time-consuming, and error-prone. As a result, the annotation process can often take hours or even days, depending on factors such as the number of samples, tissue heterogeneity, the number of spots covered by the tissue, and the level of detail required in the annotations. Annotations can be broad—such as distinguishing between tumour and non-tumour areas—or more detailed, such as identifying heterogeneous spots with mixed content, referred to as mixed spots (e.g. tumor&stroma spots in Figure 1). In particular, mixed spots can help delineate region boundaries, which are important for understanding key biological processes. For instance,

cell communication at the tumour-stroma interface plays a vital role in tumour growth and progression [32]. Therefore, identifying and annotating those spots is crucial yet highly time-consuming, as it requires examining the composition of the tissue within each spot in the anatomical boundary region. This manual and labour-intensive process of spot-level annotation significantly limits the number of samples that can be thoroughly correlated with the pathologist's input. Furthermore, the variability in annotation complexity among samples and experiments poses a major challenge for developing universal automated solutions. Currently, there are no existing solutions that fully address this issue.

To tackle these challenges, we present ActiveVisium—the first active learning-based framework, to our knowledge, that offers case- and pathologist-specific support for manual annotations in ST datasets. ActiveVisium identifies the most informative spots for pathologist manual labelling by utilizing morphological and, optionally, molecular features from tissue samples while automatically annotating the remaining spots. This framework significantly reduces the annotation workload for pathologists, resulting in a more efficient, scalable, and consistent annotation process.

2 Related work

A standard procedure in ST data analysis involves grouping spots based on shared transcriptomic profiles, morphological features, or spatial proximity. This grouping helps identify functional regions within the tissue and uncovers the *biological identity* of each spot [33], linking them to specific spatial domains or cellular niches [23,33]. Nevertheless, the biological interpretation of these groups remains a downstream step, often relying on marker genes or differential gene expression analysis, usually cross-referenced with manual pathologist annotations [23,33]. However, the time-consuming nature and complexity of manual annotation tasks pose a significant challenge to the number of samples that can be thoroughly analysed. While regional annotations (e.g., assigning a single label to a large tissue area covering several dozen spots) offer a seemingly straightforward way to reduce such manual effort, this approach often fails to account for critical spatial heterogeneity. This is particularly true for small, low-represented, or transitional regions, such as mixed or boundary spots common in technologies like 10x Visium. Therefore, despite its time commitment, individual, spot-by-spot expert annotation remains the most reliable approach for precise characterisation.

Similar bottlenecks in obtaining high-quality manual annotations also arise in the digital pathology (DP) field. Despite significant progress in Artificial Intelligence (AI)-based solutions[20], these methods still heavily depend on (manual) expert annotations, which are both resource-intensive and time-consuming to produce [28]. Consequently, numerous workflows have been developed to alleviate annotation demands, including methods to accelerate manual labelling [14,26,7].

Active learning, a paradigm aimed at maximising performance with minimal labelled data, is widely adopted in DP to minimize annotation effort. These approaches, used for tasks like cell classification [25] and whole-slide image

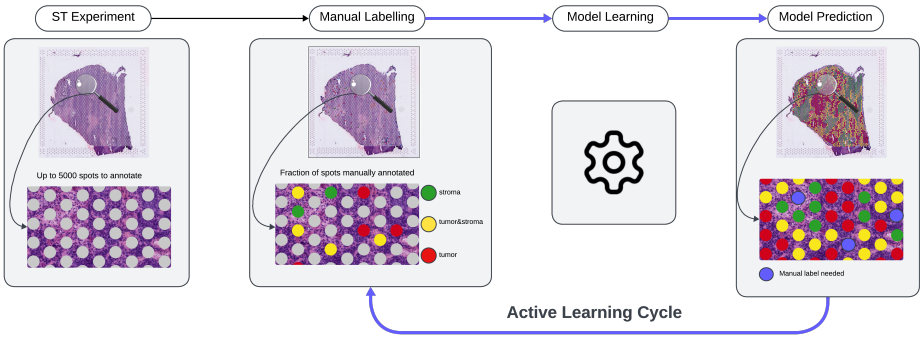

Fig. 2: The ActiveVisium framework leverages active learning to streamline spot-level annotations in ST experiments. Pathologists begin by annotating an initial subset of spots, which is used to train a model for predicting annotations on the remaining spots. Subsequently, additional spots are selected for annotation, enabling pathologists to review and refine the model's predictions. Once these new annotations are incorporated, the model is retrained with the updated dataset. This process continues until the expert is satisfied with the model's predictions or until correcting the model's potential errors requires less effort than annotating a new set of suggested spots.

annotation [16], iteratively select the most informative data points for labelling. To improve learning efficiency, many active learning techniques leverage pre-trained models, frequently initialized with ImageNet weights [9,12]. The recent emergence of foundation models in DP [2,29] is further enhancing these approaches through the development of integrated active learning frameworks [5].

The existence of a wide range of methods designed to automate the labelling process in DP underscores a crucial point: for large datasets requiring expert annotation, it is often more practical and efficient for experts to review and validate model predictions rather than manually label every individual data point [25,11]. However, such strategies have yet to be effectively translated to ST, in part due to the novelty of the field and the unique challenges posed by spot-level annotations. This motivated the development of the ActiveVisium framework, which leverages foundational models and active learning strategies to minimize the number of spots requiring manual annotation. This approach significantly reduces the workload for experts, addressing a challenge that, to the best of our knowledge, has been largely unexplored in the literature.

3 Methods

ActiveVisium is a framework that leverages active learning to optimize and accelerate manual spot-level annotations in 10x Visium ST experiments. It starts with a small set of manually annotated spots and iteratively selects additional

spots for annotation, progressively improving predictions for the remaining spots. Figure 2 presents an overview of the workflow.

In ST technologies such as 10x Visium, WSI is co-registered with the capture area containing gene expression capture spots. Let I_{WSI} represent a WSI obtained as part of such ST experiment, where the positions of gene expression capture spots are mapped onto the image. We define S_{all} as the set of tuples:

$$S_{\text{all}} = \{(x_i, g_i) \mid x_i \in \mathbb{R}^{H \times W \times 3}, g_i \in \mathbb{R}^n, i \in \{1, \ldots, N\}\}$$

where: $x_i \in \mathbb{R}^{H \times W \times 3}$ represents an image patch extracted from I_{WSI} corresponding to a capture spot i, and $g_i \in \mathbb{R}^n$ is the gene expression vector for the same spot, with n denoting the number of detected genes and N the total number of capture spots covered by tissue. The spatial dimensions $H \times W$ correspond to the pixel area covered by a capture spot (in standard 10x Visium experiments, this corresponds to a circle with a diameter of 55 μm).

Furthermore, let $S_{\text{ann}} \subset S_{\text{all}}$ represent the subset consisting of manually annotated spots provided by the pathologist ($|S_{\text{ann}}| \ll |S_{\text{all}}|$). Starting from the initial set of annotated spots $S_{\text{ann\_init}}$, the active learning pipeline is established to accelerate the process of obtaining annotations for the remaining spots in the following way:

1. **Model Learning:** The model is trained using the available annotations S_{ann} (initially $S_{\text{ann}} = S_{\text{ann\_init}}$) to predict labels for all remaining spots (see 3.1 for details).
2. **Data Acquisition Strategy:** Predictions are generated for all spots in $S_{\text{all}} \setminus S_{\text{ann}}$. Using a predefined active learning strategy, an additional set $S_{\text{to\_ann}} \subseteq S_{\text{all}} \setminus S_{\text{ann}}$ of M spots is selected for expert annotation.
3. **Expert Annotation:** The pathologist reviews the model's predictions for spots in $S_{\text{all}} \setminus (S_{\text{ann}} \cup S_{\text{to\_ann}})$ and assigns labels to the newly selected spots in $S_{\text{to\_ann}}$. The set of annotated spots is then updated as $S_{\text{ann}} = S_{\text{ann}} \cup S_{\text{to\_ann}}$.
4. **Iteration:** Steps 1–3 are repeated iteratively until the model achieves accurate predictions across S_{all}, or until correcting misclassified samples requires more effort than annotating a new set $S_{\text{to\_ann}}$, as determined by an expert.

3.1 Model training

Let $f_{\text{morph}} : \mathbb{R}^{H \times W \times 3} \to \mathbb{R}^{d_m}$ and $f_{\text{ge}} : \mathbb{R}^n \to \mathbb{R}^{d_g}$ be feature extractors mapping image patches and gene expression profiles to d_m- and d_g-dimensional embeddings, respectively. A classifier $\phi : \mathbb{R}^d \to \mathbb{R}^K$ projects embeddings into K classes, where K is specified by a pathologist.

We consider two settings - an unimodal setting where $d = d_m$ and only morphological features are used and, a multimodal setting where a fusion layer $h : \mathbb{R}^{d_m + d_g} \to \mathbb{R}^d$ combines outputs from f_{morph} and f_{ge}, providing the fused representation $h(f_{\text{morph}}(x), f_{\text{ge}}(x))$ as input to ϕ.

Morphological features are extracted using pre-trained DP foundational models. Each spot $x \in S_{\text{all}}$ is mapped to the feature representation of the foundational

model, which is stored to accelerate training. This representation is then processed through a Multi-Layer Perceptron (MLP) to obtain the morphological feature representation $f_{\text{morph}}(x)$. Similarly, gene expression features are processed by the gene expression feature extractor f_{ge}, which maps them into a latent space using an MLP. To ensure consistent feature representation across all capture spots within the sample, we identify the top 1,000 highly variable genes (HVGs) across the sample. The normalized and log-transformed expression levels of these HVGs are used as features for each spot and subsequently mapped into the latent space via an MLP. Nevertheless, given that the field of foundational models in DP and ST is actively evolving [27,18], ActiveVisium is designed to support the seamless integration of state-of-the-art and emerging models. This flexibility extends to the classifier ϕ, which is implemented as a configurable stack of MLPs, allowing for adaptability to various classification tasks.

Our model leverages the UNI framework [2] for morphological feature extraction. The classifier includes a single hidden layer with 128 neurons. Both gene expression and morphology branches use projection heads with 128 neurons and LeakyReLU activation. The fusion layer integrates these features via a normalization layer, an MLP with 256 neurons, LeakyReLU activation, and a dropout layer.

During each active learning iteration, the model is trained for 50 epochs, selecting the best-performing model based on validation loss, following standard practice in active learning literature [10]. The initial set $S_{\text{ann\_init}}$ is determined using k-means clustering in the morphological representation space of all spots, where $n_{clusters} = |S_{init}|$. In our experiments, we set the size of the initial annotation set $|S_{init}| = 55$. To ensure comprehensive class representation, in cases where the initial k-means clustering fails to encompass all classes (a scenario often encountered in datasets with highly imbalanced class distributions, such as the kidney tissue samples used in this study), the initial dataset S_{init} is augmented by randomly selecting and incorporating one spot from the annotation pool for each unrepresented class.

In real-world applications of ActiveVisium, using a predefined validation set is impractical, as annotating data solely for validation during active learning is not feasible. Consequently, this setting is used only to report experimental results. In practical applications, the training strategy adapts based on the size of S_{ann}, assuming no available validation set. For small S_{ann}, the model is trained for 50 epochs (configurable), after which the final model is retained for evaluation. With larger S_{ann}, the annotated data is split into training and validation sets during each model training, saving the model based on validation performance.

In the initial iteration, the model is initialized with random weights. Subsequent iterations reuse the model from the previous iteration as the starting point. Weighted categorical cross-entropy loss is employed to address data imbalance. The class weights are dynamically recalculated in each active learning cycle as the inverse of class frequencies.

3.2 Data Acquisition Strategy

To select M spots for manual annotation, a hybrid least-confidence and diversity-based sampling approach is implemented. At iteration i, the model predicts labels for non-annotated spots $S_i = S_{\text{all}} \setminus S_{\text{ann}}$, and each spot $x_j \in S_i$ is assigned an uncertainty score:

$$\text{score}(x_j) = \left[1 - P_{\Theta_i}(y_j^* \mid x_j)\right] \times \frac{K}{K-1}$$

where y_j^* is the highest softmax output, Θ_i are model parameters at iteration i, and K is the number of classes. This score reflects the model's uncertainty about its most confident prediction for each spot, with higher scores indicating greater uncertainty.

While active learning methods vary widely [22], least-confidence uncertainty sampling is chosen as the default due to its simplicity and effectiveness across datasets [5,28]. Nevertheless, the framework remains flexible regarding its active learning strategy. Recent evaluations confirm that incorporating diversity into uncertainty-based selection enhances performance and can outperform more complex strategies [5,3]. Therefore, diversity is incorporated using a k-means clustering approach: the top 5% most uncertain spots are grouped into M clusters within a feature space defined by morphological representations (or a combined feature space in a multimodal setting). From each cluster, the spot closest to the centroid is selected for annotation. As a baseline, we also include random sampling, in which M spots are chosen randomly from $S_{\text{all}} \setminus S_{\text{ann}}$, with each spot having an equal probability of selection, irrespective of uncertainty score.

3.3 Expert annotation

Expert annotations are conducted using Loupe Browser, the standard software for exploring outputs from 10x Visium experiments. Loupe provides an intuitive graphical interface that enables pathologists to interact with ST data easily. Given that pathologists are already familiar with using Loupe for manual spot annotation, we opted to integrate ActiveVisium with it to maintain this workflow. Pathologists annotate selected spots in Loupe and export the results as CSV files. Likewise, ActiveVisium generates predictions and selects spots for annotation also in CSV format, ensuring seamless import into Loupe for further review and refinement.

3.4 Datassets

ActiveVisium framework was evaluated on a diverse set of 10x Visium ST datasets, encompassing human and mouse samples from various tissues and pathological conditions, including breast and colorectal cancer (CRC) (human), as well as healthy kidney tissue (human and mouse). Breast cancer and kidney samples were FFPE-preserved samples, while CRC samples were FF-preserved. Experienced pathologists manually annotated each dataset at the spot level, determining the

Table 1: Dataset Summary

Specimen	Tissue	Pres.	Reference	Spots	Classes	Ann. Time
Human	Colorectal (Cancer)	FF	SN048_A121573_Rep1	2,750	8	∼8h
			SN048_A121573_Rep2	2,906	7	∼8h
			SN123_A595688_Rep1	1,394	11	∼8h
	Breast (Cancer)	FFPE	–	4,992	11	∼12h
	Kidney (Healthy)	FFPE	–	5,928	11	∼12h
Mouse	Kidney (Healthy)	FFPE	–	3,124	14	∼3h

number of classes based on tissue morphology. Table 1 presents a concise overview of the datasets used, while the Appendix provides detailed dataset descriptions, including references for each dataset.

The annotation process was highly time-intensive, with complexity varying across samples. Individual samples required between 3 to 12 hours of continuous annotation time, with a typical case taking 8 hours of continuous work per sample. This substantial time investment translates to multiple days needed to complete annotations across all datasets. These annotations serve as the ground truth for evaluation purposes in this study. Each dataset is divided into a training set (annotation pool), comprising 90% of the total spots covered by a tissue, and a test set, containing the remaining 10%. Additionally, 10% of the training set is set aside as a validation set to monitor training progress. The split is performed to maintain the class distribution, ensuring a proportional representation of all classes across the data splits.

4 Results

In this section, we showcase the effectiveness of the ActiveVisium framework across various tasks: simulated annotation experiments in both unimodal and multimodal contexts (Section 4.1), cross-sample annotation transfer illustrated through a CRC case study (Section 4.2), enhancement of annotation consistency (Section 4.3), and significant time savings in the annotation process (Section 4.4). Furthermore, we provide practical guidelines for the optimal utilization of ActiveVisium (Section 4.5).

4.1 Evaluating ActiveVisium: Simulated Annotation Experiments

All datasets used in this study are entirely manually annotated by a trained pathologist. To simulate the active learning process, we initially considered

all spots in each fully annotated dataset as unlabeled. In each active learning iteration, ActiveVisium selects a subset of spots for annotation, which are then incorporated into the training data for the next iteration.

Experiments are conducted in both unimodal (morphology only) and multimodal (morphology and gene expression) settings, with three independent runs of ActiveVisium performed for each dataset within each setting. In each active learning iteration, we fixed $M = 55$ spots to be chosen for annotation, selected based on the given experimental strategy, and this process is repeated for 10 rounds. The evaluation metrics include the average weighted F-score and the percentage of misclassified spots (along with standard deviations), compared to manual annotations that are treated as the ground truth.

Following the evaluation protocol outlined in Zhang et al. (2023) [31], Figure 3 presents active learning performance on the annotation pool, which aligns closely with the purpose of ActiveVisium—annotating a whole sample using a limited amount of provided data. To ensure a comprehensive evaluation, results on held-out test sets are included in the Appendix. Additionally, fully supervised model performances (both unimodal and multimodal) are reported to estimate the upper-bound performance.

The results obtained across various datasets suggest that using AI-based assistance for annotations is beneficial. In the early stages of training, both active learning and random sampling show promising trends in reducing misclassified samples and increasing the weighted F-score. However, active learning strategies consistently outperform random sampling, demonstrating the advantages of guided annotation over random approaches. The most significant changes are observed in the initial iterations, emphasizing the importance of following the model's suggestions for annotation in the early steps. Nonetheless, the quantitative results should be interpreted with caution, taking into account the limitations and variability associated with manual annotations (see more in Section 4.3).

The impact of multimodal approaches varies among samples and does not uniformly improve performance (see Figure 3). For instance, while the breast cancer sample shows substantial improvement, the human kidney sample exhibits only minimal benefit. This inconsistency likely arises from the degree of alignment between gene expression profiles and pathologist annotations, as morphology and gene expression can capture different biological aspects of tissue. To quantify this alignment, we first performed Louvain clustering at multiple resolutions-a hyperparameter that determines the total number of clusters- on the highly variable genes in the gene expression space. We then assessed the correspondence between the resulting clusters and the manual pathologist annotations using the Adjusted Rand Index (ARI) [1]. The breast cancer sample achieved the best ARI of 0.49 (resolution 0.7), notably higher compared to the human kidney sample had the best ARI of 0.17 (resolution 0.3). Therefore, it is not surprising that the incorporation of gene expression data led to significant performance gains in the breast cancer dataset, whereas the multimodal approach offered only marginal

[1] An ARI of 1 indicates a perfect match, while an ARI close to 0 suggests the agreement is no better than random.

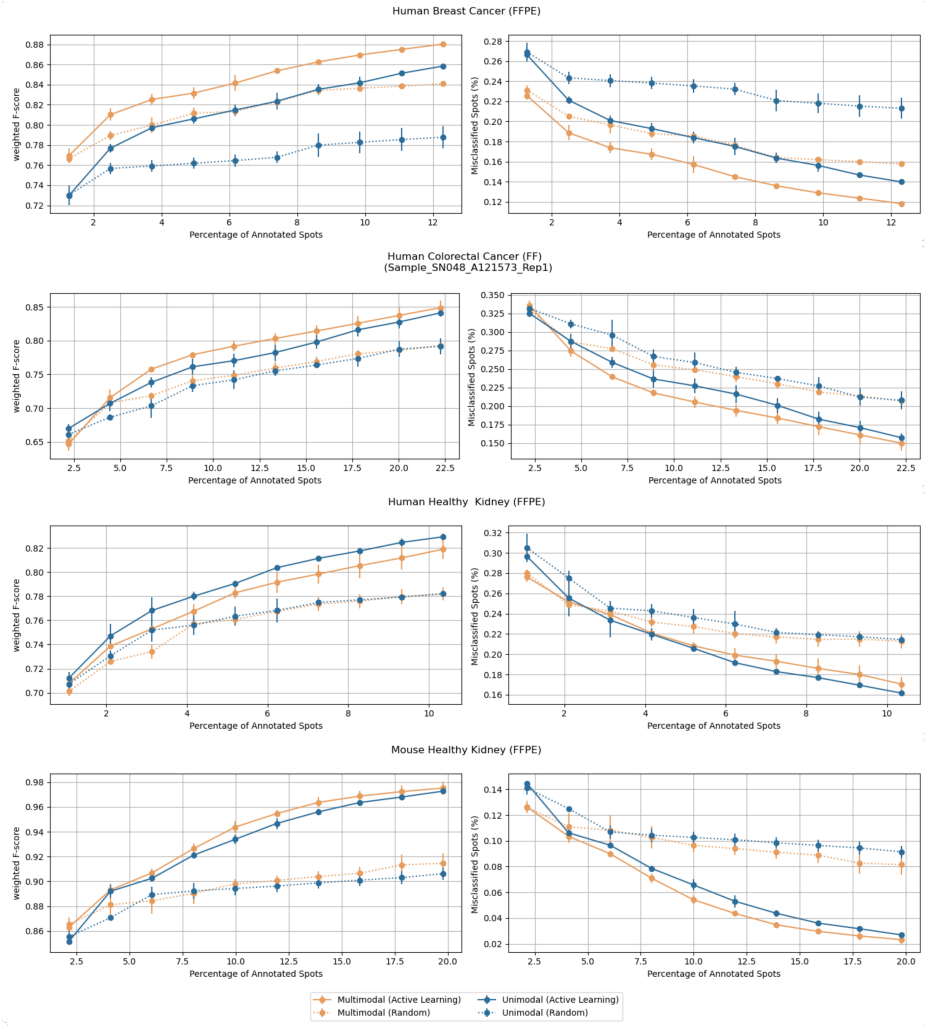

Fig. 3: Performance comparison of ActiveVisium across all evaluated datasets, including multimodal and unimodal settings. Each approach is assessed against its random sampling.

benefits in the human kidney sample. Additional ARI results across different Louvain clustering resolutions and samples are provided in the Appendix.

4.2 Cross-Sample Annotation Transfer: A Colorectal Cancer Case Study

To ensure statistically robust biological conclusions, ST studies typically comprise multiple samples from a certain tissue or disease type. Some experiments also include replicate samples from the same subject, obtained from consecutive tissue sections. In this context, our goal is to evaluate the generalizability of a model trained on a single sample by assessing its performance in two scenarios: (i) a replicate from a consecutive section of the same patient and (ii) a morphologically and pathologically similar sample from a different patient.

We evaluated the transfer learning ability of ActiveVisium on human FF resection samples from a CRC study [24]. Specifically, we trained a model in an active learning setting on Sample_SN048_A121573_Rep1 and subsequently applied it to its replicate, Sample_SN048_A121573_Rep2 and the sample from a different patient, Sample_SN123_A595688_Rep1. Manual annotations, obtained from the original study [24] served as ground truth. Since annotations were independently performed per sample, some classes present in the original sample were not available in the others. To ensure consistency in evaluation, we standardized the labels by merging similar classes and excluding non-corresponding ones (see Appendix).

In a zero-shot setting, we evaluated inter-observer variability by comparing the annotations provided by ActiveVisium— which has no annotated spots in this sample— to those manually annotated by a pathologist. In the unimodal setting, the average inter-observer agreement for the replicate sample was 0.59(0.01), while for the sample coming from a different patient, it was 0.52(0.02), indicating moderate agreement [13]. This level of concordance translates to a substantial proportion of correctly classified spots. For instance, on average, 89% and 82% of tumour spots are correctly classified in the replicate and in the sample from a different patient, respectively.

For multimodal annotation transfer, we first integrated and batch-corrected gene expression among samples with Harmony [8]. Then, highly variably genes were identified in the integrated set, and the active learning multimodal model is trained using the sample Sample_SN048_A121573_Rep1. Then, we applied it to the replicate and the sample from another patient. This approach yields inter-observer agreements of 0.061(0.00) for the replicate and 0.53(0.05) for the other sample, indicating substantial and moderate agreement, respectively. The multimodal approach is particularly beneficial for enhancing the detection rate of spots with mixed composition. For instance, when transferring to another sample, the percentage of correctly classified spots covering both tumour and stroma increased from 41.91% in an unimodal setting to 66.5% in a multimodal setting. These results are not surprising, as transcriptomic heterogeneity is more evident in mixed spots, where genes specific to various anatomical regions come

together in different proportions based on their composition. In contrast, it is more challenging to identify them based on morphological features only.

The ActiveVisium approach has the potential to greatly reduce annotation time by transferring labels from samples with similar morphological features. This is especially advantageous for studies involving multiple samples with similar features. However, pathologists still need to review and correct annotations, which can include leveraging initial predictions from transfer learning to initiate the active learning process.

4.3 ActiveVisium Improves Annotation Consistency

Consistency in annotations is crucial for the accuracy and reliability of ST data analysis. To assess whether ActiveVisium enhances this consistency, we evaluated its performance on an FFPE human kidney sample, where discrepancies between its predictions and expert annotations were most pronounced. Given the subjectivity of manual annotations, we investigated whether these misclassifications resulted from noisy manual labels rather than from model errors.

The human kidney sample consists of 5,928 annotated spots categorized into 11 classes. The initial manual annotation process took approximately 12 hours over several days, making it difficult to establish consistent labelling criteria, particularly in heterogeneous regions. Three independent ActiveVisium runs are conducted, each comprising 10 active learning cycles. The model's predictions were then compared to the original manual labels. We found 606 instances where all three models, after completing the 10th iteration of active learning, consistently classified these instances differently from the original manual annotations. A detailed breakdown of the misclassified instances by category is included in the Appendix. The pathologist who performed the initial manual annotations reviewed these spot annotations 20 weeks later, comparing the model predictions with the original annotations.

Upon investigation, the pathologist changed the annotations for 330 spots (54. 46%) - 307 accepted the model prediction, while 23 received new annotations (disagreement between model prediction, but also with the original annotation). This highlights inconsistencies in the original annotations and the challenge of maintaining uniform criteria over time. As expected, most of the changes occurred in heterogeneous areas (e.g., tubule vs. tubule-interstitium regions). While these findings confirm spot-level annotation inconsistencies, this experimental setting may be subject to confirmation bias [4] - experts might question their original annotations when faced with conflicting model predictions. Nevertheless, these results suggest that ActiveVisium enhances annotation consistency by promoting uniform criteria throughout the process. This is particularly evident in certain classes, such as glomeruli, where the model successfully identified and corrected annotations that were clearly overlooked or misclassified in the original manual process (see Appendix).

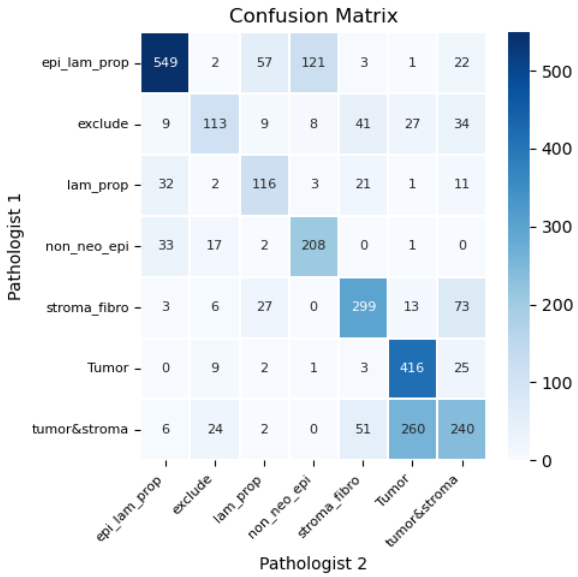

Fig. 4: Confusion matrix comparing spot-level annotations between Pathologist 1 (manual annotation) and Pathologist 2 (ActiveVisium-assisted annotation) for Sample_SN048_A121573_Rep2.

4.4 Accelerating Annotation: Time Savings with ActiveVisium

This section quantifies the time savings achieved by ActiveVisium compared to fully manual annotation using Sample_SN048_A121573_Rep2 (2906 spots, 7 classes). Two pathologists were asked to annotate the sample—one did it entirely by hand, while the other used ActiveVisium. Pathologist 1, who performed the manual annotation at the spot level, estimated that the task took approximately 8 hours of uninterrupted work. In contrast, Pathologist 2, after a brief orientation and practice with ActiveVisium, annotated the same sample in just 1 hour and 15 minutes, achieving a moderate inter-observer agreement of 0.6 with Pathologist 1. The ActiveVisium workflow involved an initial annotation of 102 spots in 17 minutes. This was followed by two active learning iterations, each annotating 55 spots in an average of 10 minutes, resulting in 211 annotated spots (7% of all spots). After each active learning iteration, Pathologist 2 spent approximately 7 minutes inspecting model predictions. Following the second iteration, Pathologist 2 assessed that the model's predictions were sufficiently accurate for the task at hand (with minor corrections needed) and that further active learning iterations were unnecessary. This correction process, applied to 42 spots, took 13 minutes and 17 seconds, ultimately resulting in 253 pathologist-provided annotations.

Final annotations occasionally differed from Pathologist 1's. These divergences were most pronounced in heterogeneous regions—like at the transition between the tumour and the stroma—where annotation criteria are inherently subjective

and vary between experts. This is illustrated in the confusion matrix in Figure 4. Despite these discrepancies, Pathologist 2 verified that this final set accurately represents their annotation style and provides a representative annotation of the sample. These differences are consistent with the inter-observer variability commonly encountered in pathology [19], further demonstrating ActiveVisium's potential as a valuable, personalised assistant that adapts to individual pathologist workflows.

4.5 Guidelines for Effective Use of ActiveVisium

Given the human-in-the-loop nature of ActiveVisium, it is reasonable to anticipate that expert interactions may deviate from theoretical expectations, and that is what we observe in practical application. For example, experts might prioritize annotating spots they find more relevant over algorithm-selected ones, particularly in early iterations when classification criteria are not yet fully established. Additionally, they may focus on correcting model errors rather than annotating new spots, which can shift model performance closer to random sampling approaches—shown in this study to be inferior to active learning strategies. To improve ActiveVisium's efficiency, we recommend establishing clear classification criteria and annotating representative spots in the first iteration. At least two active learning iterations should be completed before prioritizing error correction, as early predictions tend to be less reliable due to limited data. Once the model stabilizes, reviewing and correcting predictions is beneficial.

5 Limitations and Conclusion

This study introduces ActiveVisium, an active learning-based framework designed to streamline manual spot annotation in 10x Visium ST experiments. It demonstrates significant potential for reducing annotation workload and improving consistency across diverse tissue types. However, we acknowledge several key limitations.

A primary challenge in evaluating active learning frameworks, such as ActiveVisium, lies in the inherent trade-off between minimizing annotation effort and the need for comprehensive performance assessment. This *validation paradox*[10] is further amplified by the relatively small dataset sizes (in terms of the number of samples available) and the difficulty in establishing a reliable ground truth due to inherent noise and inter- and intra-observer variability in pathologists' annotations. Consequently, while a traditional train-test-validation split was performed for consistency with the literature, we focus on performance within the annotation pool as a more relevant indicator of the model's utility.

The choice of the annotation tool presents another limitation. While we utilized the Loupe Browser for its familiarity and accessibility, its design is not optimal for active learning annotation tasks. Therefore, more specialized applications should be developed. This aspect was not explored at this stage of our research, as we focused on a proof-of-concept study using an established tool

which pathologists were already familiar with. Developing a dedicated tool was beyond the scope of this study. However, we envision integrating ActiveVisium into a broader workflow in the future.

Beyond the limitations discussed above, several possibilities exist for future enhancement of ActiveVisium. Given the dynamic shifts in data distribution inherent to active learning, adaptive hyperparameter tuning [15] and more sophisticated regularization techniques are expected to improve learning efficiency and robustness. Furthermore, integrating multi-scale learning and leveraging the spatial context of spots are promising strategies. Finally, given the increasing adoption of Visium HD with its significantly higher resolution ($2\mu m$ bins compared to $55\mu m$ spots in standard Visium), future work should prioritize adapting ActiveVisium to this platform, as the manual annotation of these high-resolution datasets is anticipated to present a substantial bottleneck

Looking ahead to the future applications of ActiveVisium, it is important to clarify its role in relation to pathologists. ActiveVisium is designed to complement, not replace, their expertise. By automating repetitive and time-intensive annotation tasks, it allows pathologists to focus on higher-level analysis and interpretation.

6 Data and Code Availability

All code and data used in this study—including manual pathologist annotations, configuration files, model checkpoints, and intermediate results—are available at 10.5281/zenodo.15625539 and github.com/jelica-vasiljevic/ActiveVisium.

Acknowledgments. This work was supported by the Roche Postdoctoral Fellowship (RPF) programme. We sincerely thank Andrew Janowczyk and Julio Saez-Rodriguez for productive scientific discussions around the topics covered in this manuscript.

Disclosure of Interests. All the authors are currently employed by F. Hoffmann-La Roche Ltd.

References

1. Arora, R., Cao, C., Kumar, M., Sinha, S., Chanda, A., McNeil, R., Samuel, D., Arora, R.K., Matthews, T.W., Chandarana, S., et al.: Spatial transcriptomics reveals distinct and conserved tumor core and edge architectures that predict survival and targeted therapy response. Nature Communications **14**(1), 5029 (2023)
2. Chen, R.J., Ding, T., Lu, M.Y., Williamson, D.F., Jaume, G., Song, A.H., Chen, B., Zhang, A., Shao, D., Shaban, M., et al.: Towards a general-purpose foundation model for computational pathology. Nature Medicine **30**(3), 850–862 (2024)
3. Doucet, P., Estermann, B., Aczel, T., Wattenhofer, R.: Bridging diversity and uncertainty in active learning with self-supervised pre-training. In: ICLR 2024 Workshop on Practical Machine Learning for Low Resource Settings. p. arXiv preprint arXiv:2403.03728 (2024)

4. Evans, T., Retzlaff, C.O., Geißler, C., Kargl, M., Plass, M., Müller, H., Kiehl, T.R., Zerbe, N., Holzinger, A.: The explainability paradox: Challenges for xai in digital pathology. Future Generation Computer Systems **133**, 281–296 (2022)
5. Gupte, S.R., Aklilu, J., Nirschl, J.J., Yeung-Levy, S.: Revisiting active learning in the era of vision foundation models. Transactions on Machine Learning Research (TMLR) (2024)
6. Hahn, K., Amberg, B., Monné Rodriguez, J.M., Verslegers, M., Kang, B., Wils, H., Saravanan, C., Bangari, D.S., Long, S.Y., Youssef, S.A., et al.: Points to consider from the estp pathology 2.0 working group: Overview on spatial omics technologies supporting drug discovery and development. Toxicologic Pathology p. 01926233241311258 (2025)
7. Koohbanani, N.A., Jahanifar, M., Tajadin, N.Z., Rajpoot, N.: Nuclick: a deep learning framework for interactive segmentation of microscopic images. Medical Image Analysis **65**, 101771 (2020)
8. Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P.r., Raychaudhuri, S.: Fast, sensitive and accurate integration of single-cell data with harmony. Nature methods **16**(12), 1289–1296 (2019)
9. Lee, S., Amgad, M., Mobadersany, P., McCormick, M., Pollack, B.P., Elfandy, H., Hussein, H., Gutman, D.A., Cooper, L.A.: Interactive classification of whole-slide imaging data for cancer researchers. Cancer research **81**(4), 1171–1177 (2021)
10. Lüth, C., Bungert, T., Klein, L., Jaeger, P.F.: Navigating the pitfalls of active learning evaluation: A systematic framework for meaningful performance assessment. In: NeurIPS 2023 (2023)
11. Lutnick, B., Manthey, D., Becker, J.U., Ginley, B., Moos, K., Zuckerman, J.E., Rodrigues, L., Gallan, A.J., Barisoni, L., Alpers, C.E., et al.: A user-friendly tool for cloud-based whole slide image segmentation with examples from renal histopathology. Communications medicine **2**(1), 105 (2022)
12. Ma, S., Du, H., Curran, K.M., Lawlor, A., Dong, R.: Adaptive curriculum query strategy for active learning in medical image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 48–57. Springer (2024)
13. McHugh, M.L.: Interrater reliability: the kappa statistic. Biochemia medica **22**(3), 276–282 (2012)
14. Miao, R., Toth, R., Zhou, Y., Madabhushi, A., Janowczyk, A.: Quick annotator: an open-source digital pathology based rapid image annotation tool. The Journal of Pathology: Clinical Research **7**(6), 542–547 (2021)
15. Munjal, P., Hayat, N., Hayat, M., Sourati, J., Khan, S.: Towards robust and reproducible active learning using neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 223–232 (2022)
16. Qiu, J., Wilm, F., Öttl, M., Schlereth, M., Liu, C., Heimann, T., Aubreville, M., Breininger, K.: Adaptive region selection for active learning in whole slide image semantic segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 90–100. Springer (2023)
17. Rao, A., Barkley, D., França, G.S., Yanai, I.: Exploring tissue architecture using spatial transcriptomics. Nature **596**(7871), 211–220 (2021)
18. Schaar, A.C., Tejada-Lapuerta, A., Palla, G., Gutgesell, R., Halle, L., Minaeva, M., Vornholz, L., Dony, L., Drummer, F., Bahrami, M., et al.: Nicheformer: a foundation model for single-cell and spatial omics. bioRxiv pp. 2024–04 (2024)
19. Smits, L.J., Vink-Börger, E., van Lijnschoten, G., Focke-Snieders, I., van der Post, R.S., Tuynman, J.B., van Grieken, N.C., Nagtegaal, I.D.: Diagnostic variability

in the histopathological assessment of advanced colorectal adenomas and early colorectal cancer in a screening population. Histopathology **80**(5), 790–798 (2022)
20. Song, A.H., Jaume, G., Williamson, D.F., Lu, M.Y., Vaidya, A., Miller, T.R., Mahmood, F.: Artificial intelligence for digital and computational pathology. Nature Reviews Bioengineering **1**(12), 930–949 (2023)
21. Ståhl, P.L., Salmén, F., Vickovic, S., Lundmark, A., Navarro, J.F., Magnusson, J., Giacomello, S., Asp, M., Westholm, J.O., Huss, M., et al.: Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. Science **353**(6294), 78–82 (2016)
22. Tharwat, A., Schenck, W.: A survey on active learning: state-of-the-art, practical challenges and research directions. Mathematics **11**(4), 820 (2023)
23. Túrós, D., Vasiljevic, J., Hahn, K., Rottenberg, S., Valdeolivas, A.: Chrysalis: decoding tissue compartments in spatial transcriptomics with archetypal analysis. Communications Biology **7**(1), 1520 (2024)
24. Valdeolivas, A., Amberg, B., Giroud, N., Richardson, M., Gálvez, E.J., Badillo, S., Julien-Laferrière, A., Túrós, D., Voith von Voithenberg, L., Wells, I., et al.: Profiling the heterogeneity of colorectal cancer consensus molecular subtypes using spatial transcriptomics. NPJ precision oncology **8**(1), 10 (2024)
25. van der Wal, D., Jhun, I., Laklouk, I., Nirschl, J., Richer, L., Rojansky, R., Theparee, T., Wheeler, J., Sander, J., Feng, F., et al.: Biological data annotation via a human-augmenting ai-based labeling system. NPJ digital medicine **4**(1), 145 (2021)
26. Walker, C., Talawalla, T., Toth, R., Ambekar, A., Rea, K., Chamian, O., Fan, F., Berezowska, S., Rottenberg, S., Madabhushi, A., et al.: Patchsorter: a high throughput deep learning digital pathology tool for object labeling. npj Digital Medicine **7**(1), 164 (2024)
27. Wang, C.X., Cui, H., Zhang, A.H., Xie, R., Goodarzi, H., Wang, B.: scgpt-spatial: Continual pretraining of single-cell foundation model for spatial transcriptomics. bioRxiv pp. 2025–02 (2025)
28. Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., Song, Z.: A comprehensive survey on deep active learning in medical image analysis. Medical Image Analysis p. 103201 (2024)
29. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature pp. 1–8 (2024)
30. Xu, Z., Wang, W., Yang, T., Li, L., Ma, X., Chen, J., Wang, J., Huang, Y., Gould, J., Lu, H., et al.: Stomicsdb: a comprehensive database for spatial transcriptomics data sharing, analysis and visualization. Nucleic acids research **52**(D1), D1053–D1061 (2024)
31. Zhang, J., Chen, Y., Canal, G., Das, A.M., Bhatt, G., Mussmann, S., Zhu, Y., Bilmes, J., Du, S.S., Jamieson, K., et al.: Labelbench: A comprehensive framework for benchmarking adaptive label-efficient learning. Journal of Data-centric Machine Learning Research (2023)
32. Zhang, W., Huang, P.: Cancer-stromal interactions: role in cell survival, metabolism and drug sensitivity. Cancer biology & therapy **11**(2), 150–156 (2011)
33. Zhou, Y., He, W., Hou, W., Zhu, Y.: Pianno: a probabilistic framework automating semantic annotation for spatial transcriptomics. Nature Communications **15**(1), 2848 (2024)