Inferring single-cell spatial gene expression with tissue morphology via explainable deep learning

Yue Zhao[1], Elaheh Alizadeh[1], Yang Liu[1], Ming Xu[2], J Matthew Mahoney[3*], Sheng Li[1,4*]

[1] The Jackson Laboratory for Genomic Medicine, Farmington, CT, USA.

[2] UConn Center on Aging, Department of Genetics and Genome Sciences, UConn Health, Farmington, CT, USA.

[3] The Jackson Laboratory for Mammalian Genetics, Bar Harbor, ME, USA.

[4] Department of Biochemistry and Molecular Medicine, Norris Comprehensive Cancer Center, University of Southern California, Los Angeles, CA (07/2024-).

*Corresponding author(s).

E-mails: Matt.Mahoney@jax.org; Sli68423@usc.edu

Abstract

The spatial arrangement of cells is vital in developmental processes and organogenesis in multicellular life forms. Deep learning models trained with spatial omics data uncover complex patterns and relationships among cells, genes, and proteins in a high-dimensional space, providing new insights into biological processes and diseases. State-of-the-art *in silico* spatial multi-cell gene expression methods using histological images of tissue stained with hematoxylin and eosin (H&E) to characterize cellular heterogeneity. These computational techniques offer the advantage of analyzing vast amounts of spatial data in a scalable and automated manner, thereby accelerating scientific discovery and enabling more precise medical diagnostics and treatments.

In this work, we developed a vision transformer (ViT) framework to map histological signatures to spatial single-cell transcriptomic signatures, named SPiRiT (Spatial Omics Prediction and Reproducibility integrated Transformer). Our framework was enhanced by integrating cross validation with model interpretation during hyper-parameter tuning. SPiRiT predicts single-cell spatial gene expression using the matched histopathological image tiles of human breast cancer and whole mouse pup, evaluated by Xenium (10x Genomics) datasets. Furthermore, ViT model interpretation reveals the high-resolution, high attention area (HAR) that the ViT model uses to predict the gene expression, including marker genes for invasive cancer cells (*FASN*), stromal cells (*POSTN*), and lymphocytes (*IL7R*). In an apple-to-apple comparison with the ST-Net Convolutional Neural Network algorithm, SPiRiT improved predictive accuracy by 40% using human breast cancer Visium (10x Genomics) dataset. Cancer biomarker gene prediction and expression level are highly consistent with the tumor region annotation. In summary, our work highlights the feasibility to infer spatial single-cell gene expression using tissue morphology in multiple-species, i.e., human and mouse, and multi-organs, i.e., mouse whole body morphology. Importantly, incorporating model interpretation and vision transformer is expected to serve as a general-purpose framework for spatial transcriptomics.

1 Introduction

Determining how cells and their states are arranged in space is crucial for grasping how tissue structure relates to its function[1]. The spatial arrangement of cells is also vital in developmental processes and organogenesis in multicellular life forms[2]. Single-cell sequencing necessitates the separation of cells from their native tissues and thus loses important spatial information[3]. The clear advantage of spatial omics technology is that it reveals the spatial organization and potential interactions of cell subpopulations within their native habitats, at a pixel scale ranging from sub-micrometers to 100s of micrometers (μm)[4-14].

State-of-the-art *in silico* spatial transcriptomics methods predict gene expression across histological images of tissue stained with hematoxylin and eosin (H&E) to characterize cellular heterogeneity at multi-cell resolution (~100 μm), Methods such as ST-Net[15], HisToGene[16], DeepSpaCE[17], Hist2ST[18], TESLA[19], BRST-Net[20], and THItoGene[21] feature deep learning predictive models trained using H&E images paired with spatial transcriptomics data using the Visium [12] platform (10x Genomics) to predict expression of hundreds of genes. However, due to Visium's 55-μm spot size for detecting gene transcripts and the 100 μm center-to-center distance between spots, each spot frequently encompasses several cells. This can hinder the ability to discern intricate tissue architecture and to analyze cell communication, e.g., pinpointing interactions between ligands and receptors. Thus, to leverage the vast amount of commonly available histopathology images for cellular heterogeneity analysis, there is an urgent need for *in silico* single-cell spatial transcriptomic methods. The advent of sub-cellular spatial transcriptomic technologies, e.g., imaging-based RNA hybridization methods (e.g., Xenium)[4-7] and spatial barcoding-based RNA sequencing methods (Seq-Scope)[9, 10], offers groundbreaking potential to decode the cellular heterogeneity of tissues including formalin-fixed paraffin-embedded (FFPE)-treated tissue sections.

Among deep learning models, vision transformer (ViT) models have been demonstrated to capture long-range spatial relationships with more robust prediction power for image classification tasks than regular convolutional neural network (CNN) models and also offer better model interpretability[22-24]. Here we leveraged Xenium (10x Genomics) datasets to train ViT models to predict single-cell resolution spatial gene expression from H&E images across multiple species and multiple mouse organs. Our ViT framework, SPiRiT, i.e., Spatial Omics Prediction and Reproducibility Integrated Transformer, maps histological signatures to spatial single-cell transcriptomic signatures. By integrating cross-validation and model interpretation[25] for explainable deep learning during hyper-parameter-tuning, SPiRiT accurately predicts single-cell spatial gene expression from histopathological images of human breast cancer and mouse whole pup tissues, evaluated using 10x Genomics Xenium datasets. Our approach demonstrated a 40% improvement in predictive accuracy over the ST-Net CNN algorithm for the human breast cancer 10x Genomics Visium dataset. These results highlight the capacity of SPiRiT for spatial single-cell gene expression inference across multiple species and organs. Importantly, incorporating explainable deep learning models via model interpretation and vision transformer is expected to contribute as a general-purpose framework for spatial transcriptomics.

2 Results

2.1 SPiRiT overview

We developed the SPiRiT framwork use training and testing datasets from single-cell resolution spatial gene expression data processed from the Xenium In Situ platform (10x Genomics) and post-Xenium paired H&E images of the *same* tissue section[6]. The fact that Xenium-assayed tissue sections still retain their histological structure *and thus can be H&E-stained and imaged* offers us the perfect opportunity to both train the model and assess its predictive performance, as we can apply it to the post-Xenium H&E image and compare its output to the Xenium data. To enable the model to be effectively applied to related tasks without training from scratch, we will take advantage of the machine learning approach of transfer learning. This approach allows reuse of a pre-trained model so it can be applied to a new problem, leveraging previously learned knowledge trained on a broad dataset for better generalization on different tasks. The overview of SPiRiT is displayed in **Figure 1**. SPiRiT uses Xenium resolution tile level histological images as input and predict biological signatures of interest. One challenge while using deep learning models to link interpret biomedical images is the inconsistency of models as the model has high variation based on the data it sees. SPiRiT is created specifically to handle this problem. SPiRiT is a Vision Transformer (ViT) model which is ensured to use consistent image patterns for prediction enhanced by Training and Validation Attention Consistency (TAVAC)[25]. The TAVAC score is used as an independent training metric that ensures the ViT uses the same image pattern for prediction consistently no matter the model have included the image in training or not. Moreover, the attention map for the image is used for interpretation, the High Attention Region (HAR) is used identify the meaningful biological morphology linked to the biological signature of interest (expression, pathway enrichment, cell types, etc.). Most existing tools focus on the prediction performance while SPiRiT focus on the robustness and consistency in the image pattern interpretation. The TAVAC score calculation is illustrated in[25]. It employs a two-stage cross-validation-like process where data folds swap roles between training and validation, and the model's attention maps are compared for the same images across these stages. The TAVAC score uses Pearson Correlation to quantify the consistency of these attention map. More details are included in the method section.

2.2 SPiRiT on Xenium breast cancer data identifies the correct cells while predicting cell type marker genes

In this section, we apply SPiRiT on a breast cancer Xenium data[26]. To generate Xenium data, it's imperative to delineate cell perimeters within the images before attributing transcripts to specific cells, mirroring the cell identification phases in single-cell transcriptomics[26]. This process involves identifying nuclei via DAPI staining, then extending these regions up to a 15 µm maximum or until encountering adjacent cell borders. In the analyzed sample, a total of 167,885 cells and 36,944,521 transcripts were identified, with an average of 166 transcripts per cell[26].

In this work we apply SPiRiT on Xenium breast cancer data set. We extract H&E images for each cell and use it to predict the 313 gene expression values for the cell. Some images extracted only contains black background and thus will be filtered out. After filtering we have 166453 cells left. All gene expression is log(x+1) transformed. The whole data sets is divided into training validation and test partition with the ratio of 3:1:1. Initial learning rate of 0.0001 is used with linear decay (0.5) every 50 epochs. Weight decay of 0.0001 is applied for regularization. The model is trained for 1000 epochs with batch size of 64. Mean square error (MSE) is used as the loss function. The loss achieves 0.17 for all three partitions. Pearson correlation between predicted and true expression is used for evaluation. The correlation of each gene on test partition is displayed in **Figure 2A** as a histogram. All 313

genes have positive correlation on test data. Furthermore, the ROC curve and AUC value support the robustness of the ViT model across all genes evaluated using the data from test partition (**Figure 2B**). Top 20 genes with the highest test correlations includes *KRT7* (**Figure 2C**). Correlation between the prediction and true expression value of the gene *KRT7* is 0.64 (**Figure 2D**). Among the 20 genes, 13 genes namely, *ERBB2*, *FOXA1*, *GATA3*, *MMP2*, *POSTN*, *ANKRD30A*, *CCND1*, *TACSTD2*, *CD9*, *CXCL12*, *SCD*, *FBLN1*, and *CDH1* are found to be related to breast cancer. The expression of the top 5 genes with highest prediction vs truth correlation is mapped back to the H&E slides (**Figure 2E** and **Figure S1**). pattern between true expression and prediction heatmap can be observed. We also checked the root square mean square error (RMSE) metrics and visualized the top 20 genes with lowest RMSE in **Figure S2** as a bar plot. But after checking the expression values in (**Figure S3**), the low RMSE is actually due to the many zero values in the expression. Thus, RMSE is not used for model evaluation here.

In order to demonstrate the interpretation power of SPIRIT. We further train SPIRIT to predict single gene expression. We fine-tuned the ViT with 10000 samples randomly selected. We split the whole data set into two partitions for TAVAC calculation. We picked the marker genes provided by [26], namely, *FASN* (invasive cancer cells), *POSTN* (stromal cells), and *IL7R* (lymphocytes). The annotation of different cell types is visualized in **Figure 2F**. And the true expression vs prediction heatmap is visualized in **Figure 2G-I**. One can see the high consistency between the true expression and prediction patterns in the first two columns, which can be quantitively measured Pearson Correlation around 0.5 for all three marker genes. We compare the Training Attention, Validation Attention together with the cell type annotations in **Figure 3** and **Figure S4**. One can see that the High Attention Region (HAR) which is highlighted in red overlaps with the corresponding marker gene cell type consistently. For instance, For *FASN* prediction model result in **Figure 3A**. the HAR consistently overlaps with the cells that are annotated as the invasive cancer cells (red in column 2). Similarly, same cases appear for *POSTN* and *IL7R* models in **Figure 3B** and **Figure S4**. The result shows that SPiRiT can be used to link biological signatures and tissue morphology and quantitively measure the confidence level of the association in human tissue.

2.3 SPiRiT on Xenium mouse data predict whole pup *in situ* gene expresiosn

To evaluate the generalization of our framework, we further trained a ViT model using the whole mouse pup in situ gene expression dataset using Xenium mouse tissue atlassing panel (379 genes), downloaded from 10x Genomics online resource (Xenium whole mouse data). The paired H&E image and Xenium dataset from the same tissue section contains over 1.3 million cells and 191.4 million high quality transcripts. The 379 genes identified by the Xenium panel are sequenced based on FFPE protocol. For each cell, H&E images were extracted and used to predict the expression values of these 379 genes. Images containing only a black background were filtered out, leaving 1.3 million cells for analysis. A subset of 160,000 cells was randomly downsampled, yielding good performance. All gene expression data was log(x+1) transformed. The data set was divided into training, validation, and test partitions in a ratio of 3:1:1. An initial learning rate of 0.0001 with linear decay (0.5) every 50 epochs was employed, along with a weight decay of 0.0001 for regularization. The model was trained for 1000 epochs with a batch size of 64, using mean square error (MSE) as the loss function. The loss achieved 0.13 across all three partitions. Pearson correlation between predicted and true expression values was used for evaluation.

The performance of our Xenium-based ViT model (input: H&E image, output: gene expression at ~4 μm) in predicting expression of all genes Xenium assayed achieved positive Pearson correlation coefficient of all 379 genes in the test partition (**Figure 4A**). The ROC curve (AUC=0.94) supports the robustness of the ViT model across all genes evaluated using the data from test partition (**Figure 4B)**. The top 20 genes with the highest

test correlation are presented in a bar plot, including *Ttn*, *Pygm*, and *Des*, in **Figure 4C**. The gene with the highest prediction versus truth correlation, *Ttn*, a smooth muscle gene marker, shows a Pearson correlation coefficient of 0.7 between the Xenium spatial gene expression measurement and ViT model prediction across the spot of the whole mouse pup body (**Figure 4D**). The *Ttn* gene provides instructions for making a large protein called Titin, which plays an important role in skeletal and cardiac (heart) muscle function [27], confirmed by **Figure 4E** – high expression of *Ttn* are observed on skeletal muscle regions. Other two genes, *Pygm*, encoding glycogen phosphorylase, and *C3*, encoding complement component 3 protein, are visualized in **Figure S5** and **5**, respectively, where their measured expression levels are mapped back to the H&E slides, revealing similar patterns between true expression and prediction heatmaps.

2.4 SPiRiT trained on spatial multi-cell transcriptome data outperform CNN model

We further validate the predictability and expandability of SPiRiT using an independent breast cancer spatial transcriptomic data and matched H&E images from the same tissue section (ST-Net dataset) [28]. The ST-Net dataset contains 30,612 spatially mapped gene expression profiles generated from tumor specimens from 23 breast cancer patient. In order to do apple-to-apple comparison to ST-Net, the input of Vision Transformer (ViT) model is the breast cancer tissue section H&E, and the output of ViT is the $log(x + 1)$ transformed gene expression from the same each voxel (spot) of image is used as and as the output. As ST-Net does, only 250 genes with the highest mean expression levels are included for the prediction. Leave-one-patient-out cross validation is performed where we use 22 patients' data for training and the left-out patient data for test. Next, we measure the correlation between prediction and real expression values on test patient. The metric ST-Net use for performance evaluation is number of genes predicted: the number of genes with positive prediction vs real correlation for more than 20 out of 23 patients. Results from ST-Net is aggregated as a bar plot in **Figure 6A**. The figure shows pretrained Vision transformer model has the best performance. Note that even the ViT with fresh start has better performance than ST-Net pretrained models. The experiment details are in method section. This case study shows that ViT has stronger prediction power than ST-Net.

Top 20 genes with highest correlations are attached in **Figure S6**. The top 10 genes with highest median correlations of all test patients are all cancer biomarkers. *GNAS*, *FASN*, *DDX5*, *ACTG1* are also among the top 5 well predicted genes by ST-Net [28]. GNAS is G Protein Subunit Alpha S. *GNAS* mutations have been found in various types of tumors, including pituitary, thyroid, and pancreatic [29]. *ACTG1* (encoding Actin Gamma 1) has been implicated in tumorigenesis and metastasis in various cancers [30]. *DDX5*, encoding DEAD-Box Helicase 5, has been implicated in various cancers, including breast and lung cancer [31]. *GNAS* gene expression pattern is consistent with the tumor annotation (**Figure 6B-C**), similar to *ATCG1* and *DDX5* (**Figure S7-8)**. Both true expression and prediction values are z-score normalized so that the mean is zero and variance is one. After encoding tumor and non-tumor labels as 1 and 0, we calculated the Pearson correlations between the ViT prediction and the tumor label. The correlations for the C1 and C2 regions are 0.623 and 0.621, respectively. This indicates that the ViT prediction is biologically meaningful. Tumor features in the histological image is learned to predict the cancer biomarker gene expression so that the prediction heatmap has a similar pattern as the tumor annotation based on a pathologist. The model is trained on other patients during the leave-one-patient-out cross validation; thus the performance is not due to overfitting on training data.

3 Methods

3.1 Vision Transformer Model architecture

The Vision Transformer (ViT) [32], a pioneering machine learning model developed by Google Brain, marks a notable departure from the conventional application of con- volutional networks in visual recognition tasks. Rather, it adopts the Transformer model, traditionally utilized in natural language processing, and applies it to image processing. In this architecture, images are deconstructed into sequences of patches, akin to the treatment of text as sequences of words or characters within the domain of language. Each image patch is linearly embedded and then interpreted as an input token, with the addition of position embeddings to preserve spatial information. Subsequently, the model is trained end-to-end through the utilization of self-attention and feed-forward networks, which empowers the architecture with the capacity to model intricate, long-range interactions amongst patches within an image. This innovative approach has yielded substantial advancements in numerous image classification tasks, hence attesting to the robustness and versatility of Transformer models beyond the scope of language comprehension.

3.1.1 Pretrained Vision Transformer Model architecture

The Vision Transformer (ViT) model, specifically the ViT-L/32 configuration, trained on the ImageNet-1k dataset, is a comprehensive and layered architecture intended for image classification tasks. It adopts the architecture principles of Transformer models, typically used in the realm of natural language processing, and adapts it to handle image data.

In the ViT-L/32 configuration, 'L' stands for 'Large', indicating a larger model variant. It comprises 24 transformer layers, each embedding 1024 nodes. An image is divided into a grid of non-overlapping patches, where each patch size is 32x32 pixels. The flattened patch is then passed through a linear projection to generate a 1024-dimensional embedding. These embeddings are processed as input tokens for the Transformer model. The model also introduces a special learnable [CLS] token at the beginning of the sequence of patch embeddings, which is later used for classification. Position embeddings are added to these image patch embeddings to incorporate spatial information, which is crucial in image understanding tasks. Each transformer layer in the architecture consists of a multi-head self-attention mechanism and a two-layer feed-forward neural network, facilitating the model's ability to learn com- plex, long-range interactions among the image patches. The final output of the model is obtained by applying a Multi-Layer Perceptron (MLP) to the [CLS] token's representation.

The MLP consists of two fully connected (FC) layers with 512 hidden nodes each. The FC layer has nonlinear activation function: Rectified Linear Unit (ReLU). Then the output is followed by a dropout layer (drop rate = 0.5). The final output is softmax layer (classification output) or a 3rd FC layer (regression output) depends on the output data type. In the leave one patient out cross validation experiment, FC layer is used since the output is the log gene expression of 250 genes.

3.1.2 Customized Vision Transformer Model architecture

Since the data size is limited, we developed a light version of Vision transformers with only one encoder module. The architecture is shown in **Figure S11**. We trained this Vision transformer with random initialization and achieved much better result than ST-Net trained from scratch model (**Figure S6**)

3.2 Leave one patient out Cross Validation on ST-Net data

In our research, the Vision Transformer models were systematically trained with the objective of predicting the expression levels of the 250 genes demonstrating the highest mean expression values. An examination of these genes revealed that a considerable number exhibited extremely low expression levels as shown in ST-net paper [28]. Con- sequently, this led to the predicament of a low signal-to-noise ratio, which significantly complicated the prediction process.

To ascertain the generalizability of our predictions to unseen samples, we adopted a rigorous leave-one-out cross-validation method. This entailed iterative training on datasets from 22 patients while predicting the gene expression of the remaining patient. For each cross-validation fold, the model was initialized utilizing pre-trained ImageNet weights, and all weights were trained employing stochastic gradient descent, with a learning rate of 0.0001 for pretrained ViT and 0.000001 for customized ViT. Pretrained ViT is trained for 25 epochs while ViT with fresh starts takes 1500 epochs. The model operated with a batch size of 128 (pretrained) and 256 (fresh start).

During the training phase, the dataset was expanded by randomly rotating each image by 0, 90, 180°, and applying a random flip vertically or horizontally 50% of the time. Center crop image augmentation is also introduced during training.

3.3 Attention interpretation

We used attention rollout [33] to generate the attention heatmap given an input image. Attention rollout is a technique that expands on the concept of attention in trans- former models, utilized to provide interpretability to these models, which are typically characterized as “black boxes”. Given the nature of attention in transformer models where each layer has its own attention distribution, simple aggregation methods may not sufficiently capture the multi-layer, multi-head attention complexity. To overcome this limitation, attention rollout computes an approximation of the aggregated attention by rolling out the attention process over multiple layers in a sequential manner. It calculates the contribution of every input token to every output token by traversing the attention mechanism from the last layer to the first. This process generates a singular attention matrix that embodies the aggregated influence of all attention heads across all layers, thus yielding an interpretable summary of the original model’s decision-making process. As such, attention rollout contributes significantly to our understanding of how deep transformer models operate and make predictions, thereby advancing the field’s interpretability efforts. A filter on attention maps is also introduced where only the top 10 percents of high attention values are used. The original attention rollout paper suggests taking the average attention accross the attention heads but it seems taking maximum value instead generates better results. The result reported in this paper uses max function to aggregate the values from attention heads.

3.4 Train And Validation Attention Consistency (TAVAC): ranking system to quantify how trustworthy your Vision Transformer model is

We propose the Train and Validation Attention Consistency (TAVAC) score. The TAVAC score is calculated by the following process: Using a similar idea as K fold cross validation, we divided the given data set D into two folds. In the first stage, we use one fold of data as a train and the other as a validation set. We subsequently trained the model on the training dataset and evaluated it on the validation dataset. In the second stage, we switch the roles of our two subsets, making the training dataset in stage 1 validation and the validation dataset in stage 1 training. Again, we trained the model on the training dataset and evaluated its performance on the validation dataset. Finally, we examined the model’s prediction power and consistency by comparing corresponding attention results for images in the stage 1 training against stage

2 validation. In each stage, we generate the attention map for each image using Attention Rollout [33]. If an image is used as training data, the generated attention map will be defined as Training Attention. If the image is used as validation data, the attention map will be defined as Validation Attention. The consistency between Train and Validation Attention is defined as Train and Validation Attention Consistency (TAVAC) score. In this study, we used Pearson Correlation to quantify the consistency. It can be extended to other consistency scores in the future.

3.5 Using TAVAC to direct SPIRIT training

Here, we introduce an innovative approach to the training of the SPIRIT deep learning model, prioritizing not just the performance metrics typically observed, such as accuracy and F-1 score, but also the consistency of features throughout the training process. To achieve this balance, we employ the TAVAC score [25] as a criterion for hyper-parameter tuning, ensuring that the patterns utilized for prediction by the Vision Transformer (ViT) are reliable and consistent. This leads to a trade-off scenario where a high TAVAC score must be carefully weighed against the model's predictive performance. Our current strategy is stringent: we dismiss any parameter settings that result in a median TAVAC score below 0.7. From the remaining models, we then select the one with the highest TAVAC score that still delivers acceptable prediction performance. This methodology underscores the importance of feature stability, aiming to enhance the interpretability of the model's decision-making process.

3.6 Benchmark datasets

We used two primary datasets for benchmarking: Xenium In Situ data of human breast cancer samples and mouse tissue samples.

For the human breast cancer data, we utilized Sample 1, Replicate 1 from the study "High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis" published previously[26]. This dataset, referred to as Human Breast Cancer (Sample 1), comprises 167,885 cells and 36,944,521 transcripts, with an average of 166 transcripts per cell. The data was processed to extract H&E images for each cell, filtering out images with only black backgrounds, resulting in 166,453 cells for analysis. Gene expression data was log(x+1) transformed. More details can be found in the publication.

For the mouse data, we used Xenium In Situ data generated using the Mouse Tissue Atlassing panel on a one-day-old mouse pup (C57BL/6 strain) obtained from AcePix Biosciences (Xenium whole mouse data). This dataset, referred to as Mouse Pup, included 1,355,849 cells and 191,407,876 high-quality transcripts, with a median of 113 transcripts per cell. The tissue section was prepared according to standard FFPE protocols.

Additionally, we analyzed the ST-Net spatial transcriptomics dataset [28], which captures the spatial distribution of messenger RNA (mRNA) sequences within tissue sections. This dataset consisted of 23 breast cancer patients, with three microscope images of H&E-stained tissue slides and the corresponding spatial transcriptomics data for each patient. Spatial transcriptomics measures RNA expression in spots with a diameter of 100 μm arranged in a grid with a center-to-center distance of 200 μm. Each spot captures several thousand mRNA sequences, represented as a 26,949-dimensional vector of non-negative integers, indicating the number of times a gene was counted. The dataset included 26,949 distinct mRNA species across all samples. Ethical regulations were complied with for experiments involving human tissue samples, and detailed protocols were followed for RNA-seq and subsequent analysis.

4 Discussion

The utilization of deep learning within the domain of spatial omics has marked a paradigm shift in the interpretation of high-dimensional biological data. Our current work leverages the potential of the Vision Transformer (ViT) to decode the complex interplay between gene expression and cellular morphology. This innovative application not only allows for the high-throughput analysis of spatially resolved transcriptomic data but also enriches our understanding of tissue microenvironments. Notably, the introduction of the Training and Validation Attention Consistency (TAVAC) score serves as a novel metric to ensure the robustness and reliability of feature interpretation by the ViT and thus enhance explainability of the deep learning model. This consistency is crucial, particularly when considering the clinical implications of such technologies. Through the SPiRiT model, we have demonstrated a remarkable enhancement in the prediction of cancer biomarkers. The model's ability to identify and correlate tumor regions within hematoxylin and eosin (H&E) stained images with high fidelity presents a substantial advancement over conventional methodologies. This enhanced predictive capability, evidenced by a 40% improvement over ST-net models, indicates the strength of our approach.

The strategic use of TAVAC in hyperparameter tuning underscores the importance we place on model interpretability. As the field of artificial intelligence continues to grapple with the "black box" nature of machine learning models, the transparency provided by TAVAC offers a refreshing contrast. Our strict selection process, which disqualifies models with a median TAVAC score below 0.7, has ensured that only models with reliable and repeatable attention patterns are advanced. This is especially pertinent in medical diagnostics, where the reproducibility of results is as important as their accuracy. SPiRiT also differs from other non-*in silico* models (e.g., xFuse or iStar)[34, 35] that need Visium data as input to predict near-single-cell transcriptome.

Our approach has further applications in personalized medicine. By mapping transcriptomic signatures to imaging signatures with high precision, the SPiRiT framework opens avenues for the development of personalized treatment plans based on unique gene expression profiles. The ability to accurately pinpoint cancer biomarkers within the heterogeneous cellular landscape of tumors can guide the delivery of targeted therapies, thus improving patient outcomes. The promising results of our model, particularly in the field of oncology, hint at its potential to become an invaluable tool in the fight against cancer.

Moreover, the successful application of SPiRiT in recognizing and quantifying gene expression patterns in human and mouse whole body tissue sections corroborate the feasibility of integrating computational models with traditional histopathological techniques. By bridging this gap, we can considerably enhance the speed and accuracy of diagnoses, and provide pathologists with a powerful adjunct tool that can assimilate vast datasets and discern patterns beyond human perception. The predictive power of SPiRiT, complemented by the interpretive insights provided by TAVAC, establishes a comprehensive framework that is poised to make significant contributions to the understanding of complex diseases, particularly in the identification and characterization of malignancies. SCHAF model [36] has been developed to map spatial single-cell gene expression for human metastatic breast cancer and small cell lung cancer, while we showed that our SPiRiT model works for multiple species, i.e., human and mouse, and across tissues, i.e., in mouse whole body.

In conclusion, the integration of Vision Transformers with spatial omics is an exemplar of interdisciplinary innovation. The capacity to not only process but also accurately interpret the vast amounts of data generated by modern transcriptomic technologies signifies a quantum leap in both the research and clinical arenas. As we continue to refine these computational models, they will undoubtedly become central to the discovery of new biomarkers, the elucidation of disease pathogenesis, and the advancement of precision medicine. The SPiRiT model, bolstered by the TAVAC scoring system, stands at the forefront of this technological

revolution, embodying the synergy of computational prowess and biological insight.

5 Conclusion

We conclude that the SPiRiT framework represents a significant advance in spatial transcriptomics, particularly in its application to cancer research. Our findings indicate that deep learning, especially models that focus on feature consistency like SPiRiT, has the potential to offer insightful predictions that align closely with histopathological evaluations. The predictive success, particularly in cancer biomarker gene expression, underscores the clinical relevance of SPiRiT. It provides a promising avenue for future research and applications in the automated analysis of histological images, supporting faster and more reliable medical diagnoses and potentially guiding therapeutic interventions. The potential of the TAVAC score as a new standard in model training and validation has been highlighted, setting a precedent for future work in the field of computational pathology.

6 Figure Legend

Figure 1. SPiRiT workflow: spot images are extracted from the whole H&E image and used as the input for the Vision transformer models. The Vision transformer will use the spot image to predict the corresponding bio-signatures for the corresponding spot. The signature can be expression levels, pathway scores, etc. TAVAC score is calculated to guide the model selection so that consistent High Attention Region (HAR) is used by ViT no matter the model has seen the data sample during training or not.

Figure 2. **SPiRiT predicts human breast cancer single-cell resolution spatial gene expression from H&E images.** (**A**) Histogram of prediction vs true expression Pearson correlation coefficients (PCC) for each gene on human Xenium test data. All 313 genes have positive correlation. Training and test data are Xenium In Situ data of the expression of 313 genes and post-Xenium paired H&E images, aligned by Fiji BigWarp (v9.0.0). (**B**) The ROC curve and AUC value support the robustness of the ViT model across all genes evaluated using the data from test partition. (**C**) Top 20 genes with highest correlation on Xenium test data. (**D**). Scatter plot with true expression vs predicted expression values on Xenium test data for gene *KRT7*. (**E**). Matched original H&E image, heatmap visualization on true expression from Xenium and SPiRiT prediction of *KRT7*. (**F**) Cell type annotation for Xenium breast cancer data. **G-I**: cell type marker genes *FASN* (invasive cancer), *POSTN* (stromal) and IL7R (lymphocytes) true expression and prediction are visualized as heatmap on whole H&E slides. High correlations are achieved on test data. The consistency is visualized as scatter plot on the third column.

Figure 3. High Attention Region (HAR) can be used to identify corresponding cell types. Each row shows one data point. The original H&E tile image is shown in the first column. In the second column, the corresponding cell type annotation is displayed. The annotation heatmap is extracted from the whole annotation map in **Figure 2E**. The Training Attention and Validation Attention are shown respectively on column 3 and 4. The HARs (red regions) when the ViT model predicting the gene expression for *FASN* and *POSTN*, overlap with the corresponding invasive tumor cells (red cells in **Figure 3A**) and stromal cells (yellow cells in **Figure 3B**), respectively. TAVAC score calculation is visualized as scatter plot between Training Attention and Validation Attention on column 5.

Figure 4. **SPiRiT predicts mouse single-cell resolution spatial gene expression from H&E images.** (**A**) Performance of the Xenium-based ViT model is demonstrated by the histogram of correlation coefficients of prediction vs true expression for each gene on mouse Xenium test data. Training and test data are Xenium In Situ data of the expression of 379 genes and post-Xenium paired H&E images, aligned by Fiji BigWarp (v9.0.0). (**B**) The ROC curve and AUC value support the robustness of the ViT model across all genes evaluated using the data from test partition. (**C**) Top 20 genes with the highest predicting performance measured by correlation between measured and predicted gene expression levels. (**D**) The scatter plot of predicted vs. experimental measurements of *Ttn* genes and across spots in test data using PCC. (**A-D**) The test data were never seen by the VIT model during training. (**E**) H&E images, the *Ttn* gene expression via Xenium measurements (middle panels) and via ViT model prediction (right panels) of mouse whole body section and zoomed-in regions.

Figure 5. SPiRiT predicts mouse single-cell resolution spatial *C3* gene expression from H&E images. H&E images, the *C3* gene expression via Xenium measurements (middle panels) and via ViT model prediction (right panels) of 1 day mouse pup whole body tissue section and zoomed-in regions.

Figure 6. SPiRiT predicts human multi-cell resolution spatial gene expression from H&E image. (**A**) The number of genes exhibit positive correlation between predicted and

measured gene expression across deep learning architectures. Vision Transformer (ViT) outperforms other methods: ST-Net, Resnet18, Alexnet, VGG11 and Densenet121. The best performance is achieved for ViT pretrained with ImageNet dataset (blue bar). The fresh start ViT achieves second best performance. **(B-C)** Cancer biomarker gene *GNAS* expression map vs prediction and tumor annotation of two tissue sections. The prediction and expression levels are highly consistent with the tumor region annotation (yellow region).

Figure S1. SPiRiT predicts human breast cancer single-cell resolution spatial gene expression. Heatmap visualization of Xenium measured true expression and SPiRiT prediction of four genes, i.e., *CCDC80*, *KRT8*, *EPCAM*, and *FOXA1*. The consistency is visualized as scatter plot on the right**.**

Figure S2. Top 20 genes with the lowest RMSE between Xenium measured gene expression and SPiRiT predicted gene expression.

Figure S3. Top 2 genes with the lowest RSEM. Heatmap visualization of Xenium measured expression and prediction of *MPO, CRHBP, AHSP, UCP1,* and *PDCD1*. The consistency is visualized as scatter plot on the right**.** Zero inflation is the reason of low RMSE.

Figure.S4. **IL7R model HAR overlap with lymphocytes cells (blue).**

Figure S5. **SPiRiT predicts mouse single-cell resolution spatial *Pygm* gene expression from H&E images.** H&E images, the gene *Pygm* expression via Xenium measurements (middle panels) and via ViT model prediction (right panels) of mouse whole body section and zoomed-in regions.

Figure S6. **Top 20 genes with highest test prediction correlation using Visium data from breast cancer specimens.**

Figure S7. Cancer biomarker gene ACTG1 expression map vs prediction and tumor annotation. The prediction and expression level are highly consistent with the tumor region annotation (yellow region).

Figure S8. **Cancer biomarker gene DDX5 expression map vs prediction and tumor annotation.** The prediction and expression level are highly consistent with the tumor region annotation (yellow region).

7 Acknowledgments

Dr. Sheng Li is supported by the following grants: R35GM133562, U01HG013175, U01CA271830. Dr. Matthew Mahoney is supported by R01GM141309. Drs. Matthew Mahoney and Sheng Li are supported by U54AG079753, as the NIH Common Fund Cellular Senescence Network Consortium JAX-Sen Mouse Tissue Mapping Center (TMC). We would like to thank Drs. Nadia Rosenthal, Ron Korstanje, Paul Robson for scientific discussion and feedback. We would also like to express our sincere gratitude to senior scientific writer Carmen Robinett, whose invaluable contributions significantly enhanced the quality of this manuscript.

8 Code and Data Availability

The code is available in github: https://github.com/LabShengLi/SPiRiT.
Xenium human breast cancer data: https://www.10xgenomics.com/products/xenium-in-situ/preview-dataset-human-breast
Xenium whole mouse data: https://www.10xgenomics.com/datasets/mouse-pup-preview-data-xenium-mouse-tissue-atlassing-panel-1-standard
ST-net data is downloaded at: https://data.mendeley.com/datasets/29ntw7sh4r/5

9 References

1. Crosetto, N., Bienko, M. & van Oudenaarden, A. Spatially resolved transcriptomics and beyond. *Nat Rev Genet* **16**, 57-66 (2015).
2. Steventon, B., Busby, L. & Arias, A.M. Establishment of the vertebrate body plan: Rethinking gastrulation through stem cell models of early embryogenesis. *Dev Cell* **56**, 2405-2418 (2021).
3. Macosko, E.Z. et al. Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202-1214 (2015).
4. Chen, K.H., Boettiger, A.N., Moffitt, J.R., Wang, S. & Zhuang, X. RNA imaging. Spatially resolved, highly multiplexed RNA profiling in single cells. *Science* **348**, aaa6090 (2015).
5. He, S. et al. High-plex imaging of RNA and proteins at subcellular resolution in fixed tissue by spatial molecular imaging. *Nat Biotechnol* **40**, 1794-1806 (2022).
6. 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).
7. Eng, C.L. et al. Transcriptome-scale super-resolved imaging in tissues by RNA seqFISH. *Nature* **568**, 235-239 (2019).
8. Zhang, M. et al. Spatially resolved cell atlas of the mouse primary motor cortex by MERFISH. *Nature* **598**, 137-143 (2021).
9. Fu, X. et al. Polony gels enable amplifiable DNA stamping and spatial transcriptomics of chronic pain. *Cell* **185**, 4621-4633 e4617 (2022).
10. Cho, C.S. et al. Microscopic examination of spatial transcriptome using Seq-Scope. *Cell* **184**, 3559-3572 e3522 (2021).
11. Zhang, D. et al. Spatial epigenome-transcriptome co-profiling of mammalian tissues. *Nature* **616**, 113-122 (2023).
12. Gracia Villacampa, E. et al. Genome-wide spatial expression profiling in formalin-fixed tissues. *Cell Genom* **1**, 100065 (2021).
13. Deng, Y. et al. Spatial-CUT&Tag: Spatially resolved chromatin modification profiling at the cellular level. *Science* **375**, 681-686 (2022).
14. Deng, Y. et al. Spatial profiling of chromatin accessibility in mouse and human tissues. *Nature* **609**, 375-383 (2022).
15. He, B. et al. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nat Biomed Eng* **4**, 827-834 (2020).
16. Pang, M., Su, K. & Li, M. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021.2011.2028.470212 (2021).
17. 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).
18. Zeng, Y. et al. Spatial transcriptomics prediction from histology jointly through Transformer and graph neural networks. *Brief Bioinform* **23** (2022).
19. Hu, J. et al. Deciphering tumor ecosystems at super resolution from spatial transcriptomics with TESLA. *Cell Syst* **14**, 404-417 e404 (2023).
20. Rahaman, M.M., Millar, E.K.A. & Meijering, E. Breast cancer histopathology image-based gene expression prediction using spatial transcriptomics data and deep learning. *Sci Rep* **13**, 13604 (2023).
21. Jia, Y., Liu, J., Chen, L., Zhao, T. & Wang, Y. THItoGene: a deep learning method for predicting spatial transcriptomics from histological images. *Brief Bioinform* **25** (2023).
22. Ding, M. et al. arXiv:2204.03645 (2022).

23. Dosovitskiy, A. et al. arXiv:2010.11929 (2020).
24. Yu, X. et al. arXiv:2309.10625 (2023).
25. Zhao, Y., Agyemang, D., Liu, Y., Mahoney, M. & Li, S. Quantifying Interpretation Reproducibility in Vision Transformer Models with TAVAC. *bioRxiv* (2024).
26. Janesick, A. et al. High resolution mapping of the tumor microenvironment using integrated single-cell, spatial and in situ analysis. *Nature Communications* **14**, 8353 (2023).
27. Gigli, M. et al. A review of the giant protein titin in clinical molecular diagnostics of cardiomyopathies. *Frontiers in cardiovascular medicine* **3**, 21 (2016).
28. He, B. et al. Integrating spatial gene expression and breast tumor morphology via deep learning. *Nature biomedical engineering* **4**, 827-834 (2020).
29. Weinstein, L.S., Liu, J., Sakamoto, A., Xie, T. & Chen, M. Minireview: GNAS: normal and abnormal functions. *Endocrinology* **145**, 5459-5464 (2004).
30. Dong, X., Han, Y., Sun, Z. & Xu, J. Actin Gamma 1, a new skin cancer pathogenic gene, identified by the biological feature-based classification. *Journal of cellular biochemistry* **119**, 1406-1419 (2018).
31. Du, C. et al. DDX5 promotes gastric cancer cell proliferation in vitro and in vivo through mTOR signaling pathway. *Scientific reports* **7**, 42876 (2017).
32. Dosovitskiy, A. et al. (2020).
33. Abnar, S. & Zuidema, W. in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (Association for Computational Linguistics, 2020).
34. Bergenstrahle, L. et al. Super-resolved spatial transcriptomics by deep data fusion. *Nat Biotechnol* **40**, 476-479 (2022).
35. Zhang, D. et al. Inferring super-resolution tissue architecture by integrating spatial transcriptomics with histology. *Nat Biotechnol* (2024).
36. Comiter, C. et al. Inference of single cell profiles from histology stains with the Single-Cell omics from Histology Analysis Framework (SCHAF). *BioRxiv*, 2023.2003. 2021.533680 (2023).

Figure 1

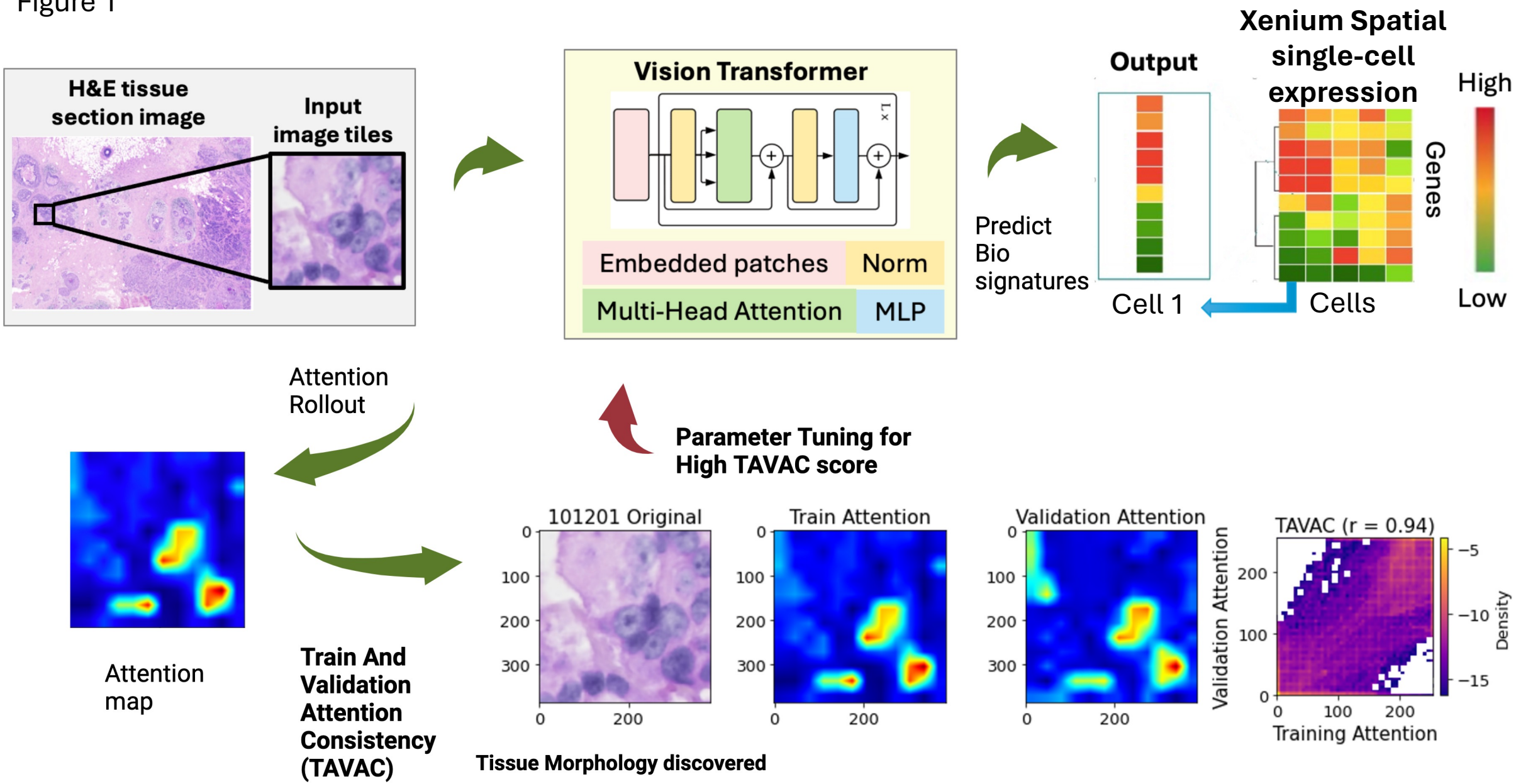

H&E tissue section image
Input image tiles
Vision Transformer
L x
Embedded patches
Norm
Multi-Head Attention
MLP
Predict Bio signatures
Output
Cell 1
Xenium Spatial single-cell expression
Genes
Cells
High
Low
Attention Rollout
Parameter Tuning for High TAVAC score
Attention map
Train And Validation Attention Consistency (TAVAC)
101201 Original
Train Attention
Validation Attention
TAVAC (r = 0.94)
Validation Attention
Training Attention
Density
Tissue Morphology discovered

Figure 2

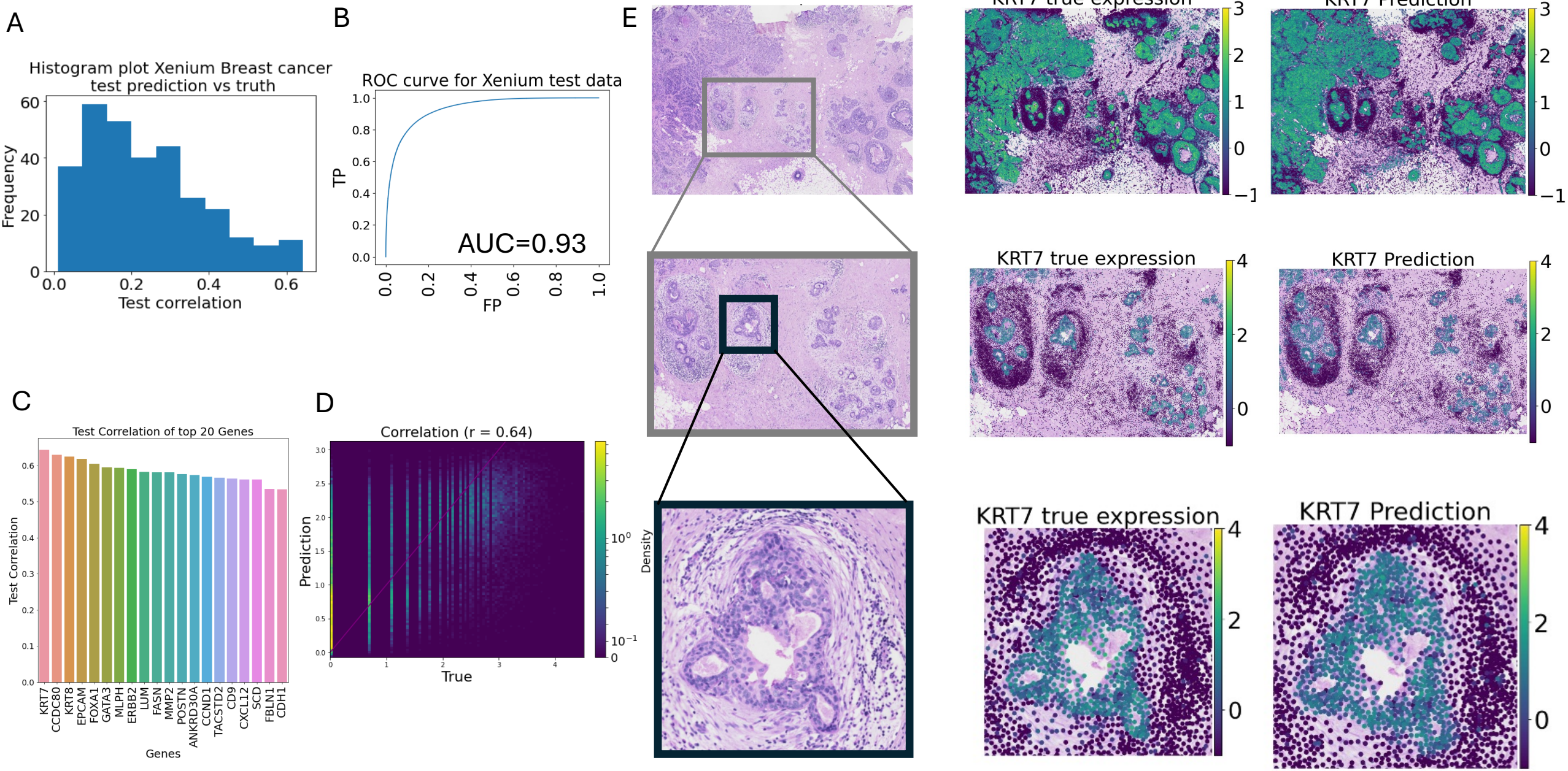

A
Histogram plot Xenium Breast cancer test prediction vs truth
Frequency
Test correlation
B
ROC curve for Xenium test data
TP
FP
AUC=0.93
C
Test Correlation of top 20 Genes
Test Correlation
Genes
KRT7
CCDC80
KRT8
EPCAM
FOXA1
GATA3
MLPH
ERBB2
LUM
FASN
MMP2
POSTN
ANKRD30A
CCND1
TACSTD2
CD9
CXCL12
SCD
FBLN1
CDH1
D
Correlation (r = 0.64)
Prediction
True
Density
E
KRT7 true expression
KRT7 Prediction
KRT7 true expression
KRT7 Prediction
KRT7 true expression
KRT7 Prediction

Figure 2

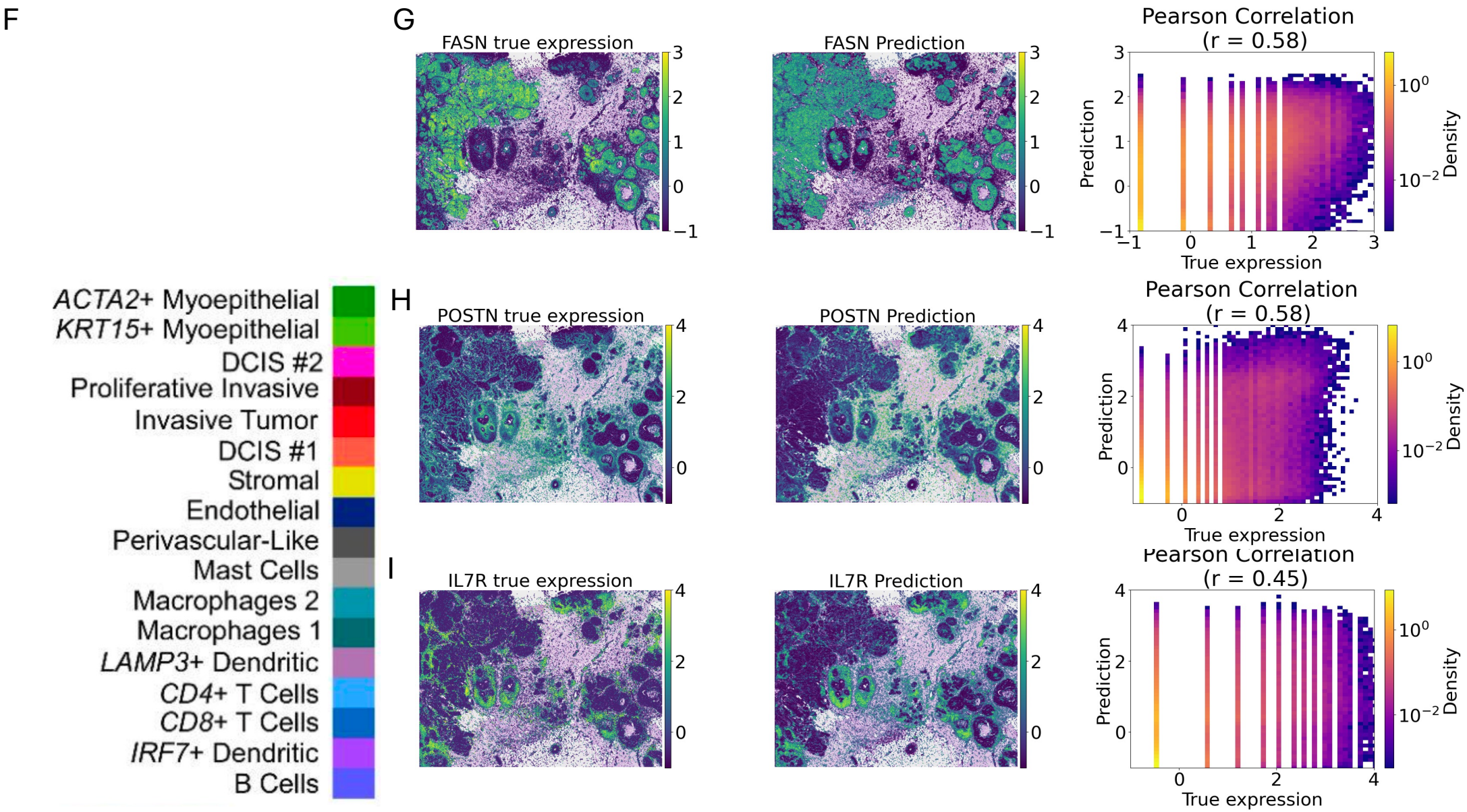

F
ACTA2+ Myoepithelial
KRT15+ Myoepithelial
DCIS #2
Proliferative Invasive
Invasive Tumor
DCIS #1
Stromal
Endothelial
Perivascular-Like
Mast Cells
Macrophages 2
Macrophages 1
LAMP3+ Dendritic
CD4+ T Cells
CD8+ T Cells
IRF7+ Dendritic
B Cells
G
FASN true expression
FASN Prediction
Pearson Correlation
(r = 0.58)
Prediction
True expression
Density
H
POSTN true expression
POSTN Prediction
Pearson Correlation
(r = 0.58)
Prediction
True expression
Density
I
IL7R true expression
IL7R Prediction
Pearson Correlation
(r = 0.45)
Prediction
True expression
Density

Figure 3

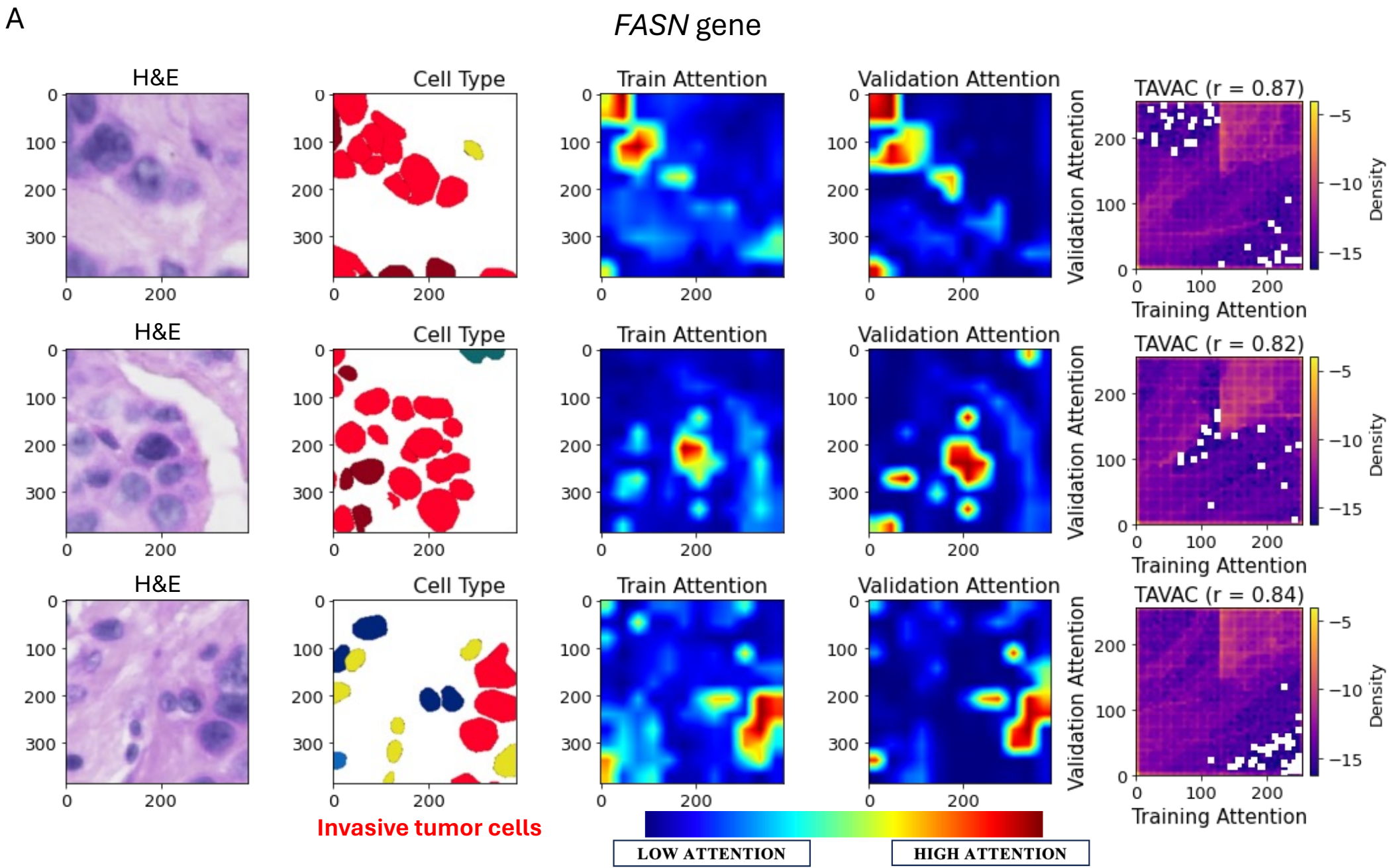

A
FASN gene
H&E
Cell Type
Train Attention
Validation Attention
TAVAC (r = 0.87)
TAVAC (r = 0.82)
TAVAC (r = 0.84)
Validation Attention
Training Attention
Density
Invasive tumor cells
LOW ATTENTION
HIGH ATTENTION

Figure 3

B

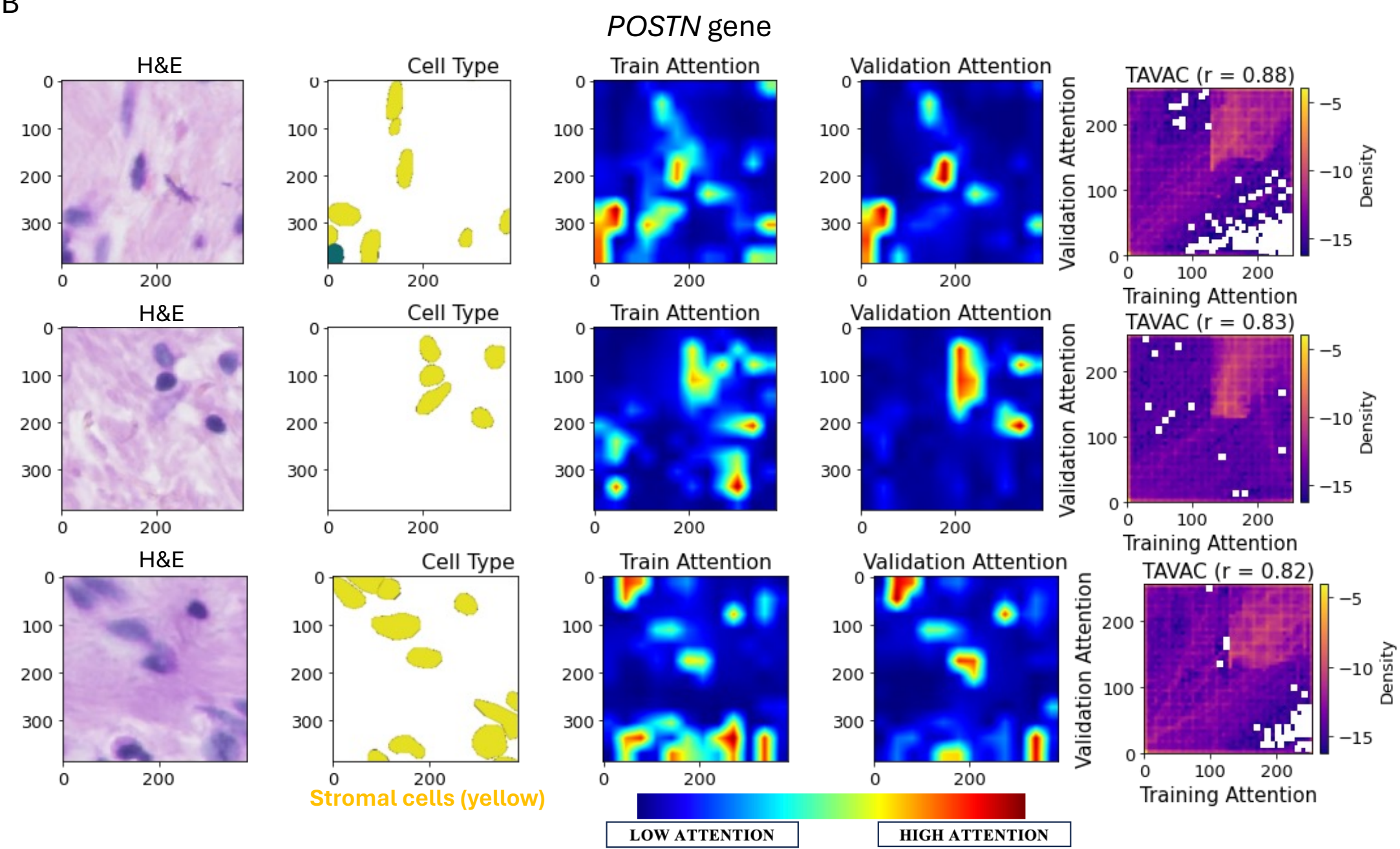
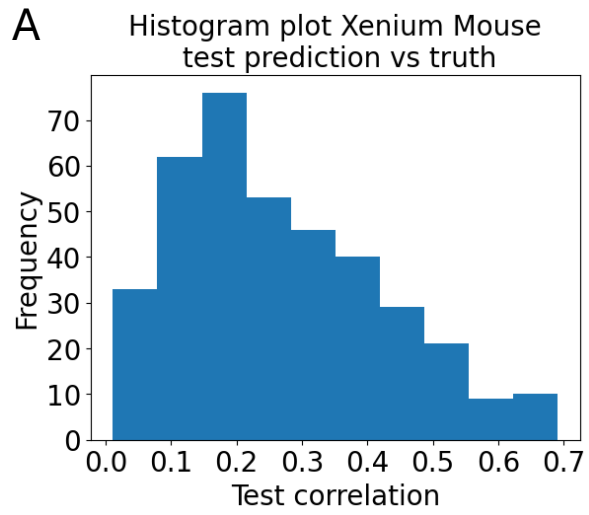
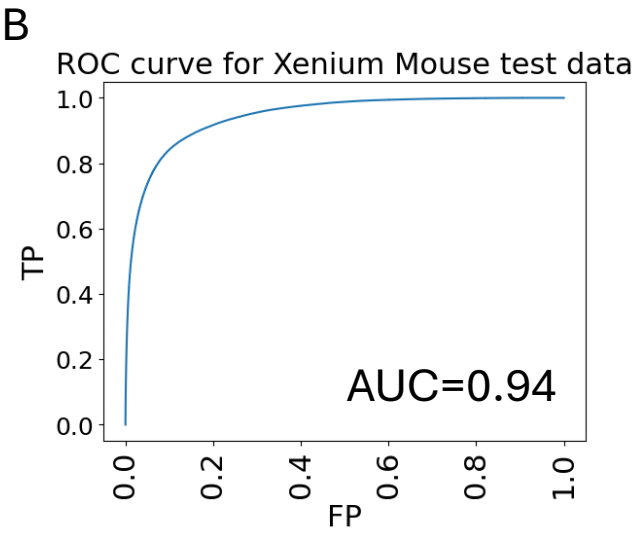
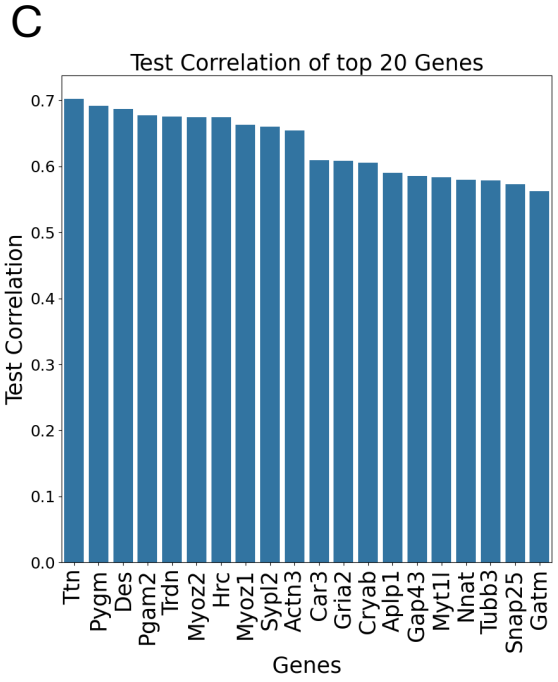
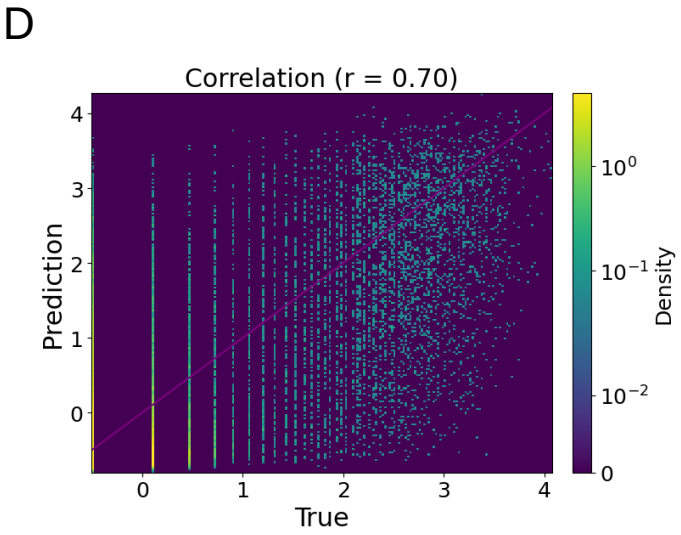

POSTN gene
H&E
Cell Type
Train Attention
Validation Attention
TAVAC (r = 0.88)
TAVAC (r = 0.83)
TAVAC (r = 0.82)
Validation Attention
Training Attention
Density
Stromal cells (yellow)
LOW ATTENTION
HIGH ATTENTION

Figure 4


A
Histogram plot Xenium Mouse test prediction vs truth
Frequency
Test correlation


B
ROC curve for Xenium Mouse test data
TP
FP
AUC=0.94


C
Test Correlation of top 20 Genes
Test Correlation
Genes
Ttn
Pygm
Des
Pgam2
Trdn
Myoz2
Hrc
Myoz1
Sypl2
Actn3
Car3
Gria2
Cryab
Aplp1
Gap43
Myt1l
Nnat
Tubb3
Snap25
Gatm


D
Correlation (r = 0.70)
Prediction
True
Density

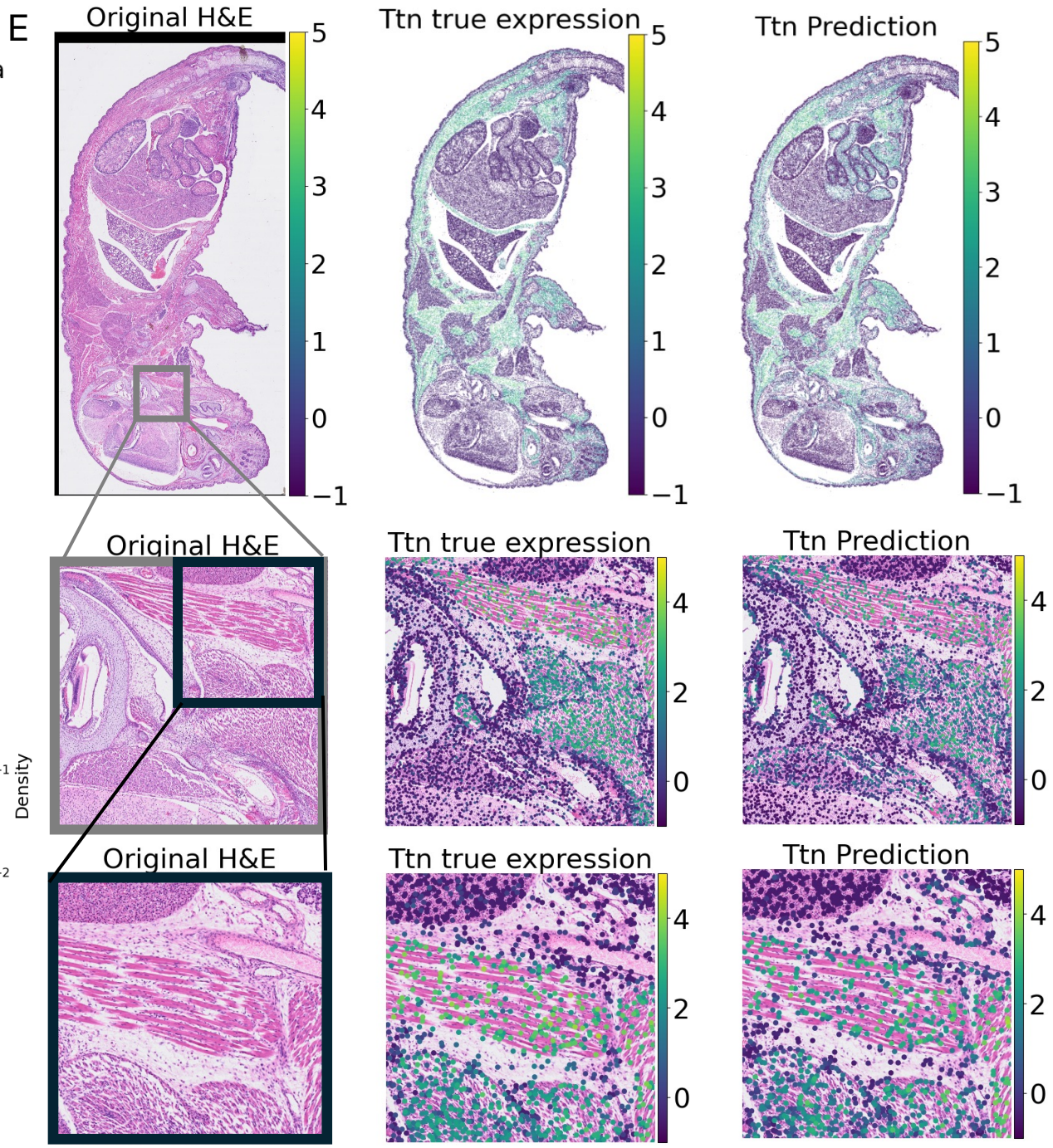

E
Original H&E
Ttn true expression
Ttn Prediction
Original H&E
Ttn true expression
Ttn Prediction
Original H&E
Ttn true expression
Ttn Prediction

Figure 5

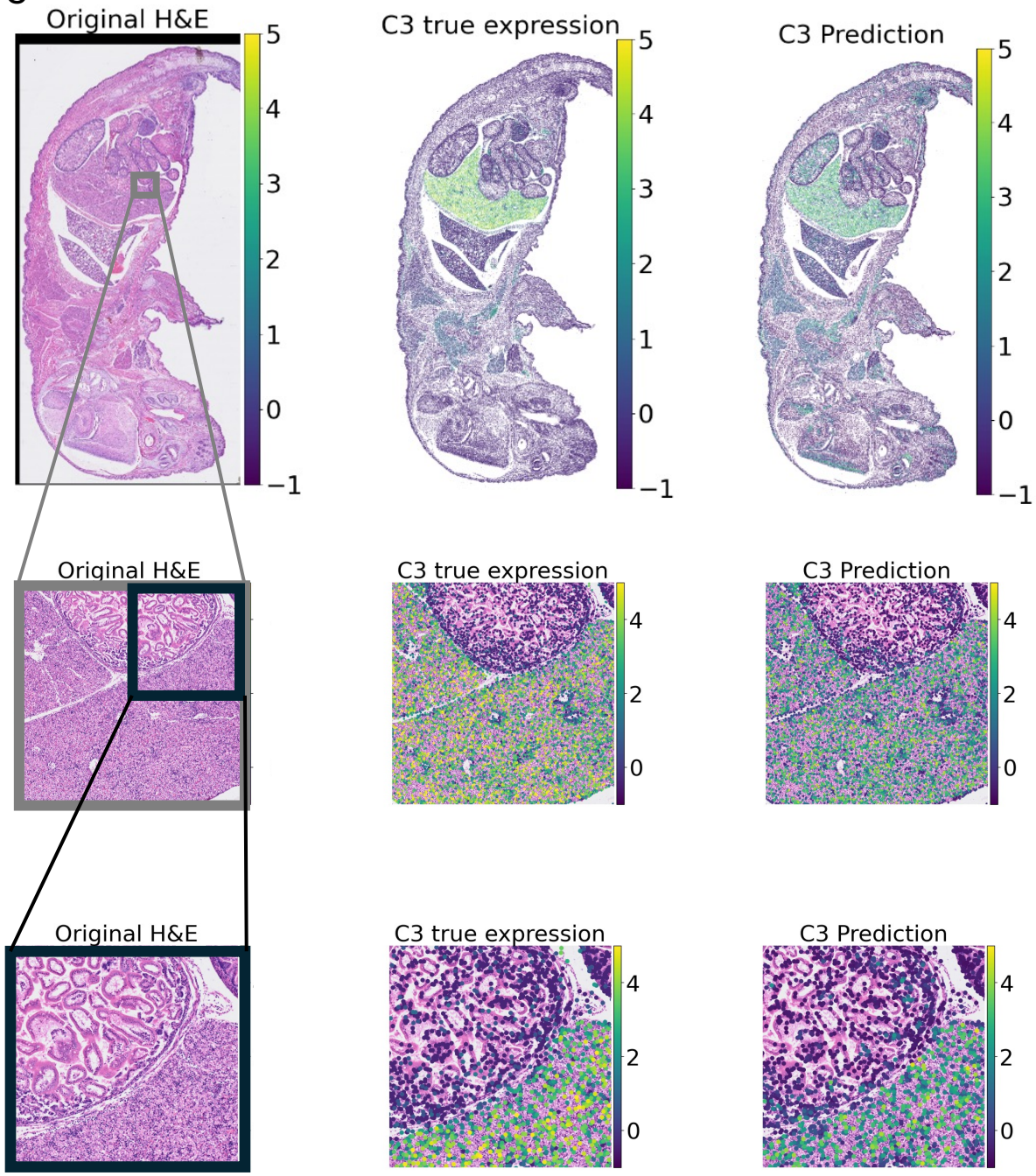
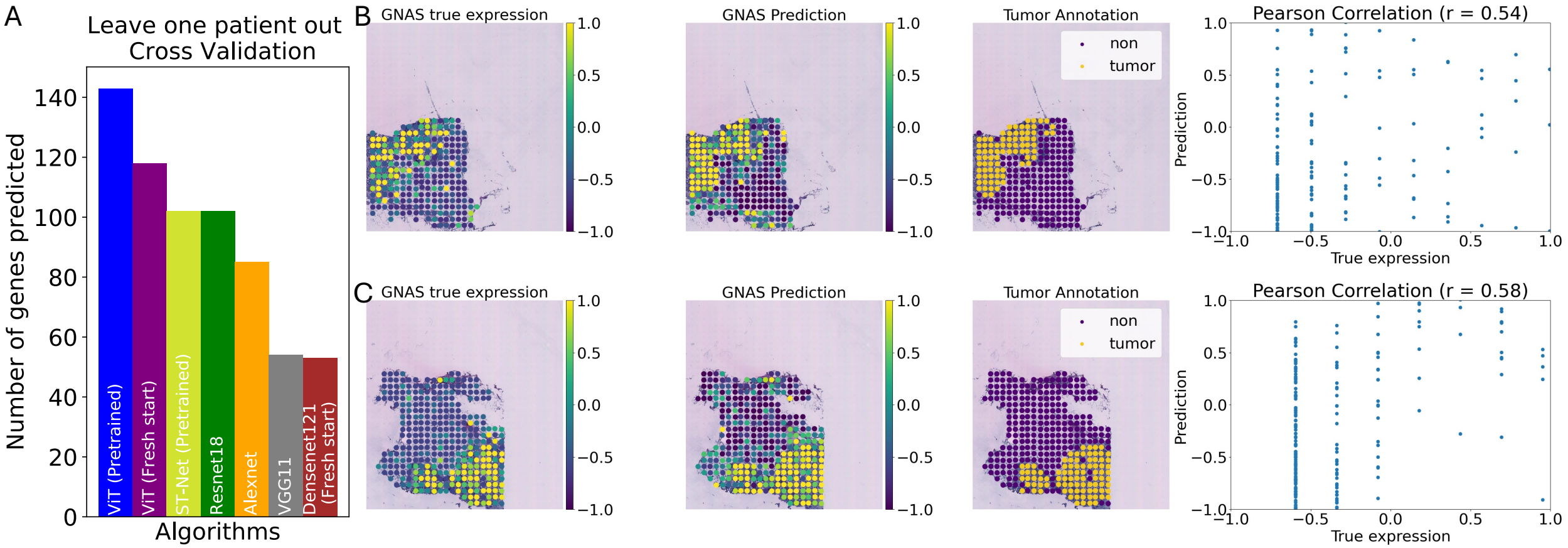

Original H&E
C3 true expression
C3 Prediction
5
4
3
2
1
0
−1
Original H&E
C3 true expression
C3 Prediction
4
2
0
Original H&E
C3 true expression
C3 Prediction
4
2
0

Figure 6


A
Leave one patient out
Cross Validation
Number of genes predicted
0
20
40
60
80
100
120
140
ViT (Pretrained)
ViT (Fresh start)
ST-Net (Pretrained)
Resnet18
Alexnet
VGG11
Densenet121
(Fresh start)
Algorithms
B
GNAS true expression
GNAS Prediction
Tumor Annotation
non
tumor
Pearson Correlation (r = 0.54)
Prediction
True expression
C
GNAS true expression
GNAS Prediction
Tumor Annotation
non
tumor
Pearson Correlation (r = 0.58)
Prediction
True expression