, 2022, pp. 1–17

doi: DOI HERE
Advance Accesss Publication Date: Day Month Year
Paper

PAPER

STimage:robust, confident and interpretable models for predicting gene markers from cancer histopathological images

Xiao Tan,[1] Onkar Mulay,[1] Samual MacDonald,[2,4] Taehyun Kim,[5] Jason Werry,[3] Peter T Simpson,[6] Fred Roosta,[2,3] Maciej Trzaskowski[1,4,7] and Quan Nguyen[1,*]

[1] Genomics and Machine Learning Lab, Institute for Molecular Bioscience, 306 Carmody Road, 4072, St Lucia, QLD, Australia, [2]ARC Training Centre for Information Resilience , The University of Queensland, 306 Carmody Road, 4072, St Lucia, QLD, Australia, [3]School of Mathmatics and Physics, The University of Queensland, 306 Carmody Road, 4072, St Lucia, QLD, Australia, [4]Max Kelsen, 39 Kennigo St, 4000, Spring Hill, QLD, Australia, [5]Pathology Queensland,Royal Brisbane & Women's Hospital, Herston, 4029, Brisbane, QLD, Australia, [6]UQ Centre for Clinical Research, Faculty of Medicine ,The University of Queensland, 71/918 RBWH Herston, 4029, Brisbane, QLD, Australia and [7]IntelMagik, West End, QLD, 4101, Australia
*Corresponding author. quan.nguyen@imb.uq.edu.au
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Spatial transcriptomic (ST) data enables us to link tissue morphological features with thousands of unseen gene expression values, opening a horizon for breakthroughs in digital pathology. Models to predict the presence/absence, high/low, or continuous expression of a gene using images as the only input have a huge potential clinical applications, but such models require improvements in accuracy, interpretability, and robustness. We developed STimage models to estimate parameters of gene expression as distributions rather than fixed data points, thereby allowing for the essential quantification of uncertainty in the predicted results. We assessed aleatoric and epistemic uncertainty of the models across a diverse range of test cases and proposed an ensemble approach to improve the model performance and trust. STimage can train prediction models for one gene marker or a panel of markers and provides important interpretability analyses at a single-cell level, and in the histopathological annotation context. Through a comprehensive benchmarking with existing models, we found that STimage is more robust to technical variation in platforms, data types, and sample types. Using images from the cancer genome atlas, we showed that STimage can be applied to non-spatial omics data. STimage also performs better than other models when only a small training dataset is available. Overall, STimage contributes an important methodological advance needed for the potential application of spatial technology in cancer digital pathology.

Key words: Neural Network, Spatial Transcriptomics

Introduction

Hematoxylin and Eosin (H&E) staining of tissue samples has been used by pathologists for more than a century. However, microscopic examination of H&E images is often time-consuming and variable [1]. Moreover, a range of molecularly distinct cell types may not be distinguishable by eyes [2]. Recent FDA approvals [3] of whole-slide imaging (WSI) for primary diagnostic use, encourage more widespread adoption of digital WSI [4]. However, the development and optimisation of automated cancer classification models requires pathologist-annotated tissue images and a huge collection of images [5, 6]. HE2RNA [7] was among the pioneers not to use pathological annotation to train a model. Instead, HE2RNA uses matched H&E images and RNA-seq data from the TCGA database for 8725 patients and 28 different cancer types, to predict gene expression at the tile level (each image is split into multiple tiles). However, due to the type of the data input, HE2RNA does not have a ground truth at the tile level, but uses the aggregated gene expression value from the whole image (all tiles in the image have the same ground truth for one gene), thus lacking power in assessing tumour heterogeneity.

Spatial transcriptomics (ST-seq) produce both imaging and sequencing information and have the unique capability of measuring over 20,000 genes without tissue dissociation. Thus preserving tissue anatomy and the microenvironment context [8]. However, spatial transcriptomics are still prohibitively expensive for clinical applications. Recently, researchers have attempted to predict spatial gene expression using deep learning methods, including STnet [9], Hist2Gene[10], a Vision Transformer in Hist2ST [11], and DeepSpace [12]. However, these methods were still limited by interpretability, robustness and performance. Crucially, clinical applications require models to be understandable to clinicians and patients for safety, trust,

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

validation and therapeutic scheduling [13]. Here, we report STimage, a deep-learning probabilistic framework with reduced uncertainty and improved robustness and interpretability, applicable even for the cases where the training datasets are limited.

Materials and methods

Datasets

The STimage workflow contains model training, prediction, interpretation, uncertainty assessment and performance evaluation steps. The training step requires spatial transcriptomics data as input with large histology images cropped to 299 x 299. Spatial coordinates encoded by spatial barcodes for each spot and gene expression values for each matched spot were used as labels to train the model. With this setup a well-trained model can be applied to histology images (non-ST data) to predict gene expression at spot level.

Our analytics involved multiple publicly available breast cancer datasets from different technologies. We used breast cancer datasets from two fresh frozen tissue samples and 1x Formalin-Fixed Paraffin-Embedded (FFPE) tissue sample, generated by 10x genomics where the RNA was measured by Visium version 1. To further test the model's generalisability, we utilised an additional dataset from six Visium slides for HER2+ breast cancer [14]. To provide thorough benchmarking of STimage, we also employed three existing methods STnet [9], His2Gene[10] and Hist2ST[11] deploying them onto another HER2+ breast cancer dataset where the RNA was measured by legacy ST technology from He et al. 2020 [9]. This dataset had 36 tissue slides from 23 patients. For indepenedent validation, we used the whole slide histological images from TCGA HER2 cohort. This way, the trained models from STimage software were applied to non-spatial data to assess their applicability in non-spatial data.

Two main designs for the trained datasets were used in this study, hence there are two sets of results and models, differing by data and the number of predicted genes. The first dataset was acquired using Visium technology (10x Genomics) and contained nine breast cancer samples. Training and validation data included two fresh-frozen samples from 10x genomics and five fresh-frozen samples from HER2+ dataset [14], with 70 % to training data, and 30 % to validation data. Test data included 1 PPFE sample from 10x genomics and one fresh-frozen samples from HER2+ dataset [14]. These test samples were also annotated by pathologists to be used for interpretability assessment. The model applied to this dataset predicts 14 cancer and immune markers and corresponds to all results, except the model benchmarking.

The second dataset was acquired using a lower resolution (legacy) ST technology and hosted 36 breast cancer samples [9]. Leave-one-out cross-validation was used. The model applied to this dataset predicted 1000 highly variable genes (HVGs), since more data were available, and was used for benchmarking STimage against STnet [9], His2Gene[10] and Hist2ST[11].

Preprocessing of image input

Technical variation between H&E images can be caused by various factors, for example, different staining procedures, imaging hardware and settings, or operators. Such technical artefacts can affect model's training and testing, negatively impacting its generalisability. In the STimage pipeline, we perform stain normalisation for each of all the images, such that the mean R, G and B channel intensities of the normalised images were similar to those of a template images, while preserving the original colour distribution patterns. STimage uses StainTool V2.1.3 to perform Vahadane [15] normalisation as the default option. In addition, the nature of the tissue sectioning, will inevitably eventuate in some tiles that contain a low tissue coverage. These tiles should be removed. STimage uses OpenCV2 for tissue masking and removes the tiles with tissue coverage lower than 70%.

STimage regression model

We utilised both sequencing data and H&E image pixel intensity data for training STimage regression models to predict spot-level gene expressions based on imaging data, only. Since the sequencing data are count based, we assessed the use of Poisson, Negative Binomial (NB), and Zero-inflated Negative Binomial (ZINB), which are distributions commonly used with regression models for this type of data. Our assessment found that the ZINB-based model did not perform as well as the NB-based model. Further, the Poisson likelihood was also deemed unsuitable for spatial sequencing data, which is typically over-dispersed and sparse. Consequently, we adopted an NB distributed likelihood $\mathbf{NB}(\mathbf{r}, \mathbf{p})$ for the STimage regression model, where $\mathbf{r} \in \mathbb{N}_+^G$ denotes the "number of successes", $\mathbf{p} \in (0,1)^G$ the "probability of each trial's success", and G is the number of (assumed) independent genes. The NB's parameters, $\mathbf{r}$ and $\mathbf{p}$, were estimated from spatial spot image $\mathbf{s}$ with a ResNet50 CNN, f_θ, with learnable pre-trained parameters θ from the ImageNet dataset, and h_ω and h_η represent parameter spaces containing weights and biases of the neural network layers connecting latent space $\hat{\mathbf{z}}$ with two neurons $\mathbf{r}$ and $\mathbf{p}$.

$$\hat{\mathbf{z}} = f_\theta(\mathbf{s})$$

$$\hat{\mathbf{r}} = \text{Softplus}(h_\omega(\hat{\mathbf{z}})) \tag{1}$$

$$\hat{\mathbf{p}} = \text{Softmax}(h_\eta(\hat{\mathbf{z}})) \tag{2}$$

STimage regression entails the pre-trained ResNet50 CNN f_θ mapping a H&E spot tile, $\mathbf{s}$, to a latent representation $\hat{\mathbf{z}} \in \mathbb{R}^{2048}$, which is then fed into separate output layers (1) and (2), estimating $\hat{\mathbf{r}}$ and $\hat{\mathbf{p}}$, respectively. Again, we assumed genes were independent and denote estimates for the g-th gene's likelihood parameters by $\hat{\mathrm{r}}_g$ and $\hat{\mathrm{p}}_g$. Therefore, the STimage regression model defined g-th (random) gene expression $X_g|\,\mathbf{s}$, as a function of H&E spot $\mathbf{s}$, by

$$X_g|\,\mathbf{s} \sim \text{NB}(\mathbf{s}\,|\,\hat{\mathrm{r}}_g, \hat{\mathrm{p}}_g) \tag{3}$$

$$\mu_g(\mathbf{s}) = \mathbb{E}[X_g|\,\mathbf{s}] = \hat{\mathrm{r}}_g(1-\hat{\mathrm{p}}_g)/\,\hat{\mathrm{p}}_g \tag{4}$$

$$\sigma_g(\mathbf{s}) = \sqrt{\mathbb{V}[X_g|\,\mathbf{s}]} = \sqrt{\hat{\mathrm{r}}_g(1-\hat{\mathrm{p}}_g)}/\,\hat{\mathrm{p}}_g, \tag{5}$$

where $\mu_g(\mathbf{s})$ in (4) is the mean prediction for the g-th gene's expression with expectation operator $\mathbb{E}[.]$, and $\sigma_g(\mathbf{s})$ in (5) is the standard deviation function with variance operator $\mathbb{V}[.]$.

Overall, the ResNet50 CNN layers are optimised to take H&E images as input to estimate each gene expression's predictive distribution with NB's parameters r_g and p_g (for all $g \in \{1, \ldots, G\}$, Eq. (3)). The ResNet50 base, f_θ, is used to perform convolution operations and spatial filtering to extract a three-dimensional feature map (i.e., tensor) from input image tiles. The pretrained f_θ of the ResNet50 includes the GlobalAveragePooling reduction operation [16] to convert the three-dimensional tensor to the feature vector $\mathbf{z}$, whereby each $\mathbf{z}$ corresponds to a single tile/spot within a H&E image.

The fully-connected output layers h_ω and h_η were fine-tuned by maximising the NB's negative log-likelihood,

$$\sum_{g=1}^{G}\sum_{i=1}^{B}\log\frac{\Gamma(\hat{r}_{g,i})\Gamma(x_{g,i}+1)}{\Gamma(\hat{r}_{g,i}+x_{g,i})}-\hat{r}_{g,i}\log(\hat{p}_{g,i})-x_{g,i}\log(1-\hat{p}_{g,i}), \tag{6}$$

with batch size B, Gamma function $\Gamma(.)$, and $x_{g,i}$ representing the observed RNA count of g-th gene (i.e., gene expression).

We trained our model on a panel of 14 marker genes which include four immune marker genes (CD74, CD24, CD63, CD81) and ten cancer genes (COX6C, TP53, PABPC1, GNAS, B2M, SPARC, HSP90AB1, TFF3, ATP1A1, FASN).

Uncertainty assessment

STimage estimates a NegativeBinomial (NB) distribution to model each gene's (random) expression $X_g \in \{0,1,\ldots\}$ conditional to H&E spot $\mathbf{s}$, i.e., $X_g|\mathbf{s} \sim \mathrm{NB}(\hat{\mathrm{r}}_g,\hat{\mathrm{p}}_g)$. A single STimage model produces the distribution $p(X_g|\mathbf{s}) = \mathrm{NB}(\mathbf{s}|\hat{\mathrm{r}},\hat{\mathrm{p}})$. An ensemble of STimage models produces a distribution $p(X_g|\mathbf{s},\Theta)$. Therefore, following the law of total variance, we can obtain two terms accounting for uncertainty

$$\mathbb{V}[X_g|\mathbf{s}] = \mathbb{V}[\underbrace{\mathbb{E}[X_g|\mathbf{s},\Theta]}_{\mu_g(\mathbf{s})}] + \mathbb{E}[\underbrace{\mathbb{V}[X_g|\mathbf{s},\Theta]}_{\sigma_g(\mathbf{s})}], \tag{7}$$

where the variance of the mean estimates $\mathbb{V}[\mu_g(\mathbf{s})]$ defines epistemic uncertainty, and the average of variances $\mathbb{E}[\sigma_g(\mathbf{s})]$ defines the aleatoric uncertainty. Aleatoric uncertainty reflects the noise inherent in the data. Epistemic uncertainty reflects the model's self (e.g. lack of information due to small sample sizes or not optimal model parameters) and can be reduced with more data and/or improved models.

Applying the formula (7), we quantify total uncertainty for the prediction of each gene in each spot. Briefly, we trained 120 STimage models separately using the same training dataset, each with an unique random seed. Corresponding values of X_g were predicted for individual spots $\mathbf{s}$ using each trained model with the model parameter Θ. We based on the 120 individual STimage models, we computed 12 ensemble models. Each STimage ensemble model was the result from averaging NB estimates $\mu_g(\mathbf{s})$ of 10 randomly pooled STimage single models. For each of the single models and ensemble models, Pearson correlation was then calculated for the prediction accuracy metric.

STimage classification model

We transformed the raw gene expression (continuous values) to z-scores and then label them as "Low" for z-score below zero and "High" for z-score above 0 where the z-score is given by:

$$z = \frac{x-\mu}{\sigma}. \tag{8}$$

For the classification model we used the pre-trained (or optionally a fine-tuning layer) ResNet50 as a feature extractor for all tiles and then applied regularised logistic regression classifier, optimizing the 'log-loss' (binary cross entropy) function. We used "elasticnet" [17] to balance between model performance and complexity, by applying both L1 and L2 penalties. The coefficient set $\boldsymbol{\beta} = (\beta_1,\ldots,\beta_n)$ as defined in Eq. (9), where $\boldsymbol{X} = (x_1,\ldots,x_n)$ are latent space features (n=2048 features). The elastic loss L_{enet} is the sum of the binary cross entropy loss (log loss) as in the Eq. (10) and the regularisation terms as in the Eq. (11). For a dataset with N spots, we have

$$\hat{y}_i = \frac{1}{1+\exp-(\beta' X+b)}, i \in \{1,\ldots,N\}, \tag{9}$$

$$L_{logistic}(\beta) = -\frac{1}{N}\sum_{i=1}^{N} y_i\log(p_{y_i}) + \sum_{i=1}^{N}(1-y_i)\log(1-p_{y_i}), \tag{10}$$

$$L_{enet}(\beta) = L_{logistic}(\beta) + \omega\left((1-\alpha)\sum_{j=1}^{2048}\beta_j^2 + \alpha\sum_{j=1}^{2048}|\beta_j|\right) \tag{11}$$

Here, α is a mixing parameter defining the relative contribution of RIDGE and LASSO to the penalty term, and for $\alpha = 1$ the ElasticNet function reduces to LASSO regularisation(logistic loss + Sum of absolute value of weights) while for $\lambda = 0$ the function reduces to RIDGE regularisation (logistic loss + Sum of squares of weights). Here, α is a regularisation parameter and by default we have set the strength of regularisation to $\alpha = 0.1$ (the smaller value, the less regularisation). The solver "SAGA" [18] was used for optimisation, which is faster for large datasets of the same scale.

Model interpretation of STimage using LIME and SHAP Models

We implement model explainability for STimage in order to highlight important features from the image that contribute towards high expression of a gene for a spot/image-tile (Figure S5). These features can be cells or tissue regions or a cancer microenvironment. We used local, model-agnostic approaches with Local interpretable model-agnostic explanations (LIME) and SHapley Additive exPlanations (SHAP). The aim was to extract the segments from the histology image tiles to visually interpret the image features that influence the prediction, in order to correlate the segments with pathological annotations. LIME and SHAP repeatedly perturb the input to measure how the output of the trained model changes. Inputs required by the LIME and SHAP model are a trained classifier/regressor, an image-tile and a gene name.

Morphological diversity among cells in the tumour is an important hallmark that provides clues for histopathological assessment. Thus an accurate and fast nuclei segmentation strategy is needed for an explainable model. We used watershed segmentation [19] for the LIME and SHAP model to define the nuclei segments to find segments corresponding to high expression of a particular gene. We performed image transformation, converting RGB images to Haematoxylin-Eosin-DAB (HED) images before applying the watershed segmentation. Below are formulas to compute LIME and SHAP values.

LIME finds a simple model (g) that can approximate our complex model f. The total loss in LIME is given by

$$\xi = \underset{g\in G}{\operatorname{argmin}}\ L(f,g,\pi_x) + \Omega(g), \tag{12}$$

where

$$L(f,g,\pi_x) = \sum_{z,z'\epsilon Z}\pi_x(f(z)-g(z'))^2 \tag{13}$$

creates a local approximation of our trained model in the neighbourhood of a given input image (x) by a simple interpretable model (g), where (G) is a set of sparse linear regression models and (π_x) denotes the local neighbourhood of the data point x. The term $L(f,g,\pi_x)$ in (12) is a weighted

sum of squared error (SSE), where the weights are given by π_x depending on the proximity of the local neighbourhood data point to the (x). This is done to make sure that the points closest to the (x) are given more weight. The second term in (12) is a regularisation of (g) by assigning zero-weight to non-important input features.

SHAP approximates Shapley values for all features (i.e., pixels) i. The inputs required for SHAP are the same as LIME, i.e., classification model f, image-tiles x and a gene. The Shapley values ϕ_i of features, indexed by i, regarding individual data points x, are (exactly) evaluated by

$$\phi_i(f,x) = \sum_{z' \subseteq x'} \frac{|z'|!(M-|z'|-1)!}{M!}[f_x(z') - f_x(z' \backslash i))]. \quad (14)$$

Here, $x' \in \{1,0\}^M$ denotes a 'simplified' input, with M many features. $z' \in x'$ is one of 2^M many possible combinations of features. $f_x(z')$ is the model output with respect to histology image x, where pixels are set to zero according to z'. The importance of feature i for combination z' is $[f_x(z') - f_x(z' \backslash i))]$ weighed by $\frac{|z'|!(M-|z'|-1)!}{M!}$. Summing the weighted importances across all subset combinations $z' \subseteq x'$ amounts to the marginal feature importance: the Shapley value. Since exact estimation of Shapley values are computationally prohibitive, we used the KernelExplainer (x' - nuclei-segments of an image) and Explainer (x - for each pixel) functions from SHAP for model interpretability.

STimage Web Application: To provide an interactive, visual interpretability option, we have developed STimage Webapp where users can select image tiles and genes of interest to visualise the important image features responsible for the expression of the genes active in those locations. The nuclei that are involved in high expression of genes will be highlighted, providing visual interpretability of the model (Figure S6). Adjustable parameters via a slider is provided for users to run optimal watershed nuclei segmentation to be used as an input for LIME and SHAP models. The LIME and SHAP scores can be visualised in the Webapp.

Benchmarking and performance assessment

We applied a leave-one-out cross-validation strategy on the 36 HER2+ ST dataset to evaluate the STimage regression model performance and benchmarked it with most similar methods such as STnet [9], HisToGene[10] and Hist2ST [11] (Figure 2f). Briefly, all models are evaluated using the same data-splitting strategy. Namely, all the models were trained on 35 samples, with the prediction of the spatial gene expression tested on the remaining sample. This way, for each method, we obtained a score for each patient. Here the score is the Pearson correlation coefficient of the predicted gene expression values against the measured gene expression values. This was used as performance metric to assess the prediction accuracy. Pearson correlation between predicted values and measured values). For generalisability, we performed the assessment for the top 1000 highly variable genes. After removing the genes that expressed in less than 10% of the total spots, 785 genes were kept for further analysis. Each gene had a score from applying a model to a patient, so for each patient, 785 scores were used to compare the models. To make the fair comparisons, we ran both the HisToGene[10] and Hist2ST [11] models using the authors' suggested model hyperparameters. However, we failed to run STnet due to technical problems in the source code, so we rewrote the STnet model according to the original paper [9]. Once we collected all predictions from all the models, we ended up with 36 models x 4 methods x 785 values for comparisons.

We also benchmarked our STimage with those methods using an out-of-distribution (OOD) test dataset using more advanced spatial transcriptomics technology Visium. In this scenario, all models were trained on two fresh frozen breast cancer tissues from 10x Genomics and tested on an FFPE tissue which we considered as an OOD sample, because there is a significant difference in the information content and the quality between these two sample types. Using the same gene filtering approach as above, we included 982 of the top 1000 HVGs in the training and prediction. Hist2ST and his2gene have a vision transformer (ViT) component as part of their neural networks, which requires loading all spot tile images and expression values from one tissue into the model in one batch. Consequently, running this model requires more computational resources (especially memory) when applied to Visium data where there are 10 times more data than in the legacy ST datasets. The training of Hist2ST and his2gene failed on a cluster with an NVIDIA SXM2 Tesla 32 GB card, necessitating to split the Visium tissue into multiple small subsets using a sliding window approach to reduce the data loading. Each small subset contained a square tissue image in size of 25x25 spots with no overlapping. Each subset was then treated as one smaller dataset and by doing this, reduced the number of spots loaded in each batch, while still retaining the spatial neighbourhood organisation. All of the small subsets from the two fresh frozen samples were trained together by Hist2ST and his2gene using their default model hyperparameters, and the two trained models were applied to predict the gene expression in the FFPE dataset. The model performance was assessed by Pearson correlation and Moran's I (Figure 2d,e).

Results and Discussion

Advances in STimage pipeline

STimage implements an important, yet often underestimated, image preprocessing step (Figure 1a). We performed stain normalisation to account for technical variation between image inputs for the model training. To remove uninformative tiles from the image inputs, we calculated tissue coverage and filtered out those with low tissue coverage. In STimage, the image tiles are simultaneously input to fine-tune the ResNet50 model (as a minibatch of, by default, 64 tiles), (Figure 1b). All the latent features (2048 features/tile) are then connected to a layer of neurons representing parameters for the negative binomial distribution of each gene, which is optimised by a log likelihood loss function. A long with the distribution estimates, we also thorouhgly quantify the model prediction uncertainty. This is an important feature, given the expected high level of variation in the model prediction. Multiple genes can be trained simultaneously at one time, with each gene corresponding to a pair of NB parameters. This way a model can learn to predict the expression of a gene panel, either as continuous or discrete values. Importantly, we add a model interpretability analysis using LIME and SHAP with a focus on cancer and immune cell markers. Several programs designed for predicting gene expression values like STnet [9], Hist2ST [11], Hist2Gene [10], DeepSpace [12], and Xfuse [20] estimate fixed point values and without uncertainty quantification.

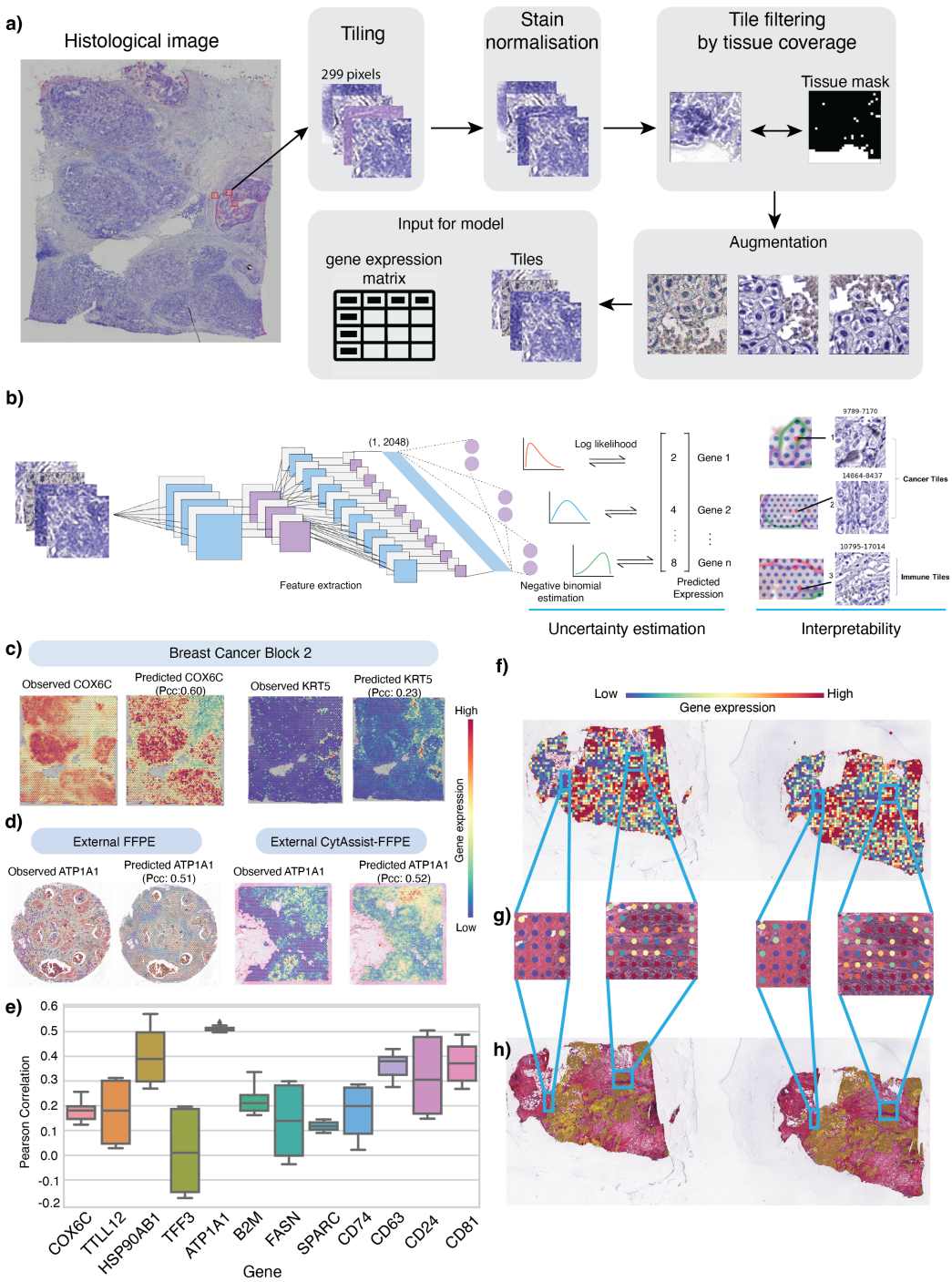

Fig. 1. The overview of the robust and interpretable STimage model. **a**, STimage image preprocessing workflow, including image tilling, stain normalisation, QC for tiles and augmentation. **b**, STimage CNN-NB interpretable regression model. STimage model consists of a convolutional neural network as an image feature extractor and negative binomial layers for estimating gene expression distribution. The negative log likelihood is used as the loss function to optimise the model parameters during the training steps. **c**, ST-measured (observed) and STimage predicted gene expression for highly abundant cancer marker COX6C and lowly abundant, but with a specific spatial pattern, keratin marker KRT5 on the test dataset. Pearson correlation (Pcc) between predicted values across all spots with the observed values, (i.e., measured by spatial transcriptomics) is shown. **d**, Model performance for external FFPE dataset from 10x Genomics. **e**, Pcc values were calculated for 12 predicted and observed cancer and immune markers for two external FFPE dataset using the ensemble model. **f,g,h** STimage regression model applied on non-spatial, whole slide image (WSI) of a HER2+ patient from the TCGA dataset. Two adjacent sections were used to compare reproducibility. **f** The WSI was cropped into small tiles in size of 299x299 pixels without overlapping. A model trained from the nine Visium samples was applied on those tiles to predict the expression of the cancer marker gene COX6C. Each spot represents a tile where red and blue colours indicate the predicted high and low COX6C expression respectively. **g** Two zoom-in views of tumour enriched area. Predicted cancer maker expression highly correlated with tumour morphology and consistent in two replicates. **h** Pathologist annotation of tumour area, (indicated by brown colour), in the two replicate tissue images.

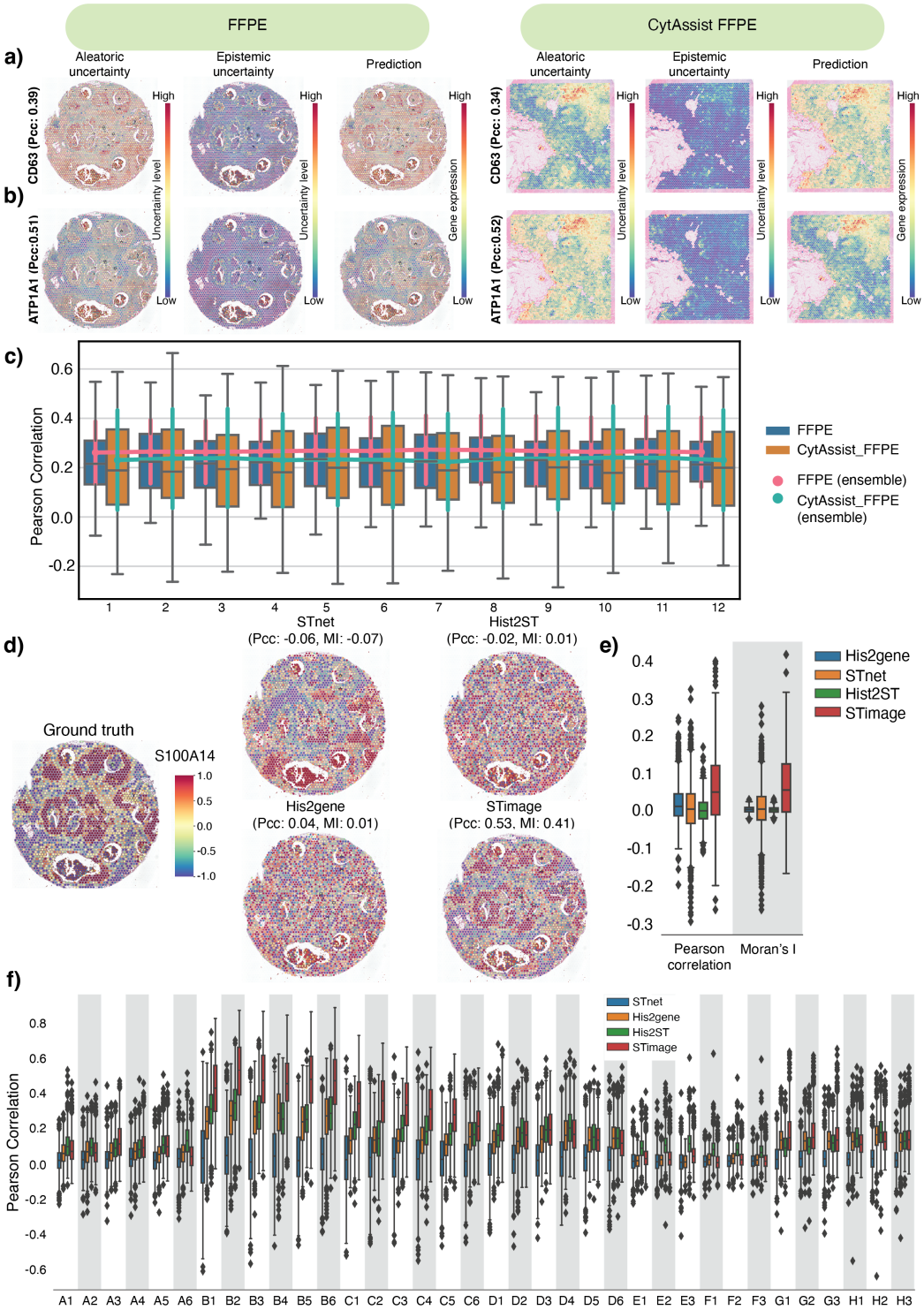

Fig. 2. Distribution-based estimation and benchmarking of STimage CNN-NB model. **a**, Uncertainty estimation and predicted gene expression for CD63 in two external FFPE datasets. The red-blue colour gradient indicates a high-low uncertainty level. **b**, The uncertainty estimation and predicted gene expression for a cancer gene ATP1A1. **c**, The performance of 12 ensemble models comparing to 10 single models in each ensemble group. The X-axis shows 12 ensemble groups where each group contains 10 runs. The y-axis is the Pearson correlation of the predicted gene expression against the ground truth. The boxes show the distribution of performance of different single model in 12 different ensemble groups and two different tissue slides (all modes have different random seed). While the lines are ensemble models using all single model in each ensemble group with error bar showing the standard deviation for different genes. **d,e** Benchmarking of STimage regression model with STnet, His2Genes and Hist2ST models using Visium data from 10x genomics. All models were trained on 2 fresh frozen breast cancer tissues and predicted on an out-of-distribution (OOD) FFPE tissue. The models were intentionally trained and tested using a small dataset and OOD samples to assess robustness. **d**, Visualisation for comparing the ground truth expression of the cancer marker S100A14 and predicted expression by different models. Pcc indicates Pearson correlation coefficient and MI indicates the Moran' I (spatail autocorrelation)**e** Pearson correlation and Moran' I score of predicted 1000 HVGs against ground truth for each model were compared. **f**, A comprehensive benchmarking with STnet, His2gene and Hist2ST model using 36 tissue slides from her2st dataset[9].

Prediction performance - higher specificity and accuracy

To assess the model specificity, we compared prediction for an abundant cancer marker (COX6C), expressed across the tissue, and a less abundant gene marker expressed specifically around the Ductal Cancer In Situ (DCIS) region (KRT5 gene). Visually, the predicted gene patterns are highly consistent with the ST measured values (ground truth), either for test datasets from Visium protocols for fresh frozen samples (Figure 1c) or the two external FFPE samples (Figure 1d). Quantitatively, using spatial autocorrelation Moran's I index, we found a significant positive correlation (Moran index with high-high or low-low scores) for COX6C across most spots (Figure S1a). This quantification is advantageous compared to models using non-spatial data, such as HE2RNA [7], because they lack per-tile gene expression measurements (bulk RNA-based data used in such models has the same value for a gene across the entire tissue), preventing the tile level performance assessment.

Assessing model performance across 14 cancer marker genes, which were trained simultaneously, we found the prediction positively correlated to the measured values for 13 of 14 genes (Figures 1e, 2e). Moreover, we obtained a higher spatial autocorrelation value than the non-spatial Pearson correlation, suggesting the prediction is spatially informative. Of note, we found that the model performance was consistently higher when applying our image normalisation strategies (Figure S2). We also thoroughly assessed the possible effect of using different resolutions for the image tile inputs. We found that the 299x299 tile, used as the default in STimage, produced the highest model prediction performance (Figure S3).

Prediction performance - more robust across diverse datasets and sample types

To assess the robustness of the approach, we increased the heterogeneity in the train and test data by including data from different protocols, laboratories, and populations. We trained a model using seven Visium datasets from fresh frozen tissues, in which two datasets were generated by 10x Genomics and five by Swarbrick's Laboratory [14]. The trained model was then tested on leave-out data, including one FFPE and one fresh frozen sample (Figure 1c,d). Visual comparison between the observed and predicted genes (COX6C) shows high consistency for both the FFPE sample and the fresh frozen sample (Figure 1d). In both tissue types, the COX6C was predicted to be high in the tumour regions, as defined by pathological annotation (Figure S1e), also consistent with the measured values. Quantitatively, we found positive Pearson correlation and Moran's spatial autocorrelation indexes between predicted and observed values for each of the 14 cancer and immune genes, suggesting that the prediction could capture the true patterns in the gene expression data.

To evaluate if the STimage could be applied to the tissue images from a non-spatial dataset, we selected two random breast cancer samples from The Cancer Genome Atlas (TCGA)(Figure 1f,g,h). To establish a ground truth for comparison, we obtained detailed pathological annotations of these two H&E images. The prediction of cancer markers was consistent across the two TCGA replicates (Figure 1f,g,h). Importantly, the prediction matches the pathological annotation, with high expression of the markers in the cancer region and low expression outside the cancer region.

Uncertainty quantification

To increase the model robustness for a broad range of diverse, unseen datasets, we trained an ensemble of STimage regression models, which have been shown to improve generalisation in the transcriptomics context [21]. In addition to improving the robustness, ensembles allow us to quantify uncertainty in terms of the data (i.e., aleatoric uncertainty), as well as uncertainty about the model itself (i.e., epistemic uncertainty). Existing methods STnet [9] and Hist2gene [10] use a CNN or a ViT as an image feature extractor. Both models use multi-layer NN to predict fixed gene expression values which did not estimate the variability in the prediction. Hist2ST [11] used a zero-inflated NB to estimate the distribution of the gene expression but does not estimate the epistemic uncertainty. To assess these new functionalities in STimage, we used the STimage model trained with the public breast cancer dataset, which consists of data from two ST spatial protocols (one for fresh frozen tissues and another for formalin-fixed tissues as described above), to test two tissues in independent, unseen cohorts (Figure 2). As an example, two representative genes, including ATP1A1 as a breast cancer marker and CD63 as an immune marker, were shown for comparisons. The aleatoric uncertainty, epistemic uncertainty, and predicted gene expression are visualised in the spatial tissue plot for these two genes (Figure2a,b). The aleatoric uncertainty (i.e. the average of predicted NB variance values across single models) was higher than the epistemic uncertainty (i.e. variance of the NB mean estimates) for both the FFPE samples. As expected, there was a tendency for both aleatoric uncertainty and epistemic uncertainties to positively correlate with the expression level. Interestingly, however, lower epistemic measures coincided with higher aleatoric measures (ATP1A1, Figure2b), hence improving one's trust on the predicted NB (i.e., the mean and variance/aleatoric). In addition to lower epistemic uncertainty, the predictive performance is much higher for the ATP1A1 gene (Pcc:0.52; Figure2b) compared to the CD63 gene (Pcc: 0.34 Figure2a), which is consistent with interpretations from epistemic uncertainty, i.e., that epistemic uncertainty explains how much a prediction can be trusted [22]. In this instance, the higher epistemic uncertainty may be partly attributed to the fact that these samples were generated from different technological platforms compared to the other samples used during training, so the data for the prediced samples was considered out-of-distribution.

An ensemble approach improves uncertainty performance

We observed variation in model prediction, even for the same model and training dataset but using different random seeds. To assess and address this issue, we analysed the distribution of the performance for individual models within 12 different groups and for two distinct tissue slides (total 120 models), annd the results are shown as in the box plots in Figure 2c. Each box shows the average Pearson correlation from independent prediction by 10 single models for each of the 12 genes. The lines represent ensemble models using average estimation from all 10 single models in each ensemble group, with error bars indicating the standard deviation for different genes (12 genes). Although there is considerable variation among individual models, their performance remains consistent across all ensemble groups. Furthermore, the ensemble models outperform the single models and demonstrate less variation across all groups (Figure 2c).

Out of distribution performance

To assess model performance on out-of-distribution (OOD) dataset, we designed an experiment where the STimage model and the three other models, including STnet, His2gene and Hist2ST, were trained on data from fresh frozen samples (Visium poly-A captured data) and predicted FFPE data (Visium probe-hybridization data). The probe-hybridization data detected more genes and more information per gene. The two datasets were derived from different tissue types and technologies and can be considered an OOD dataset (Figure 2e,f). For a small training dataset, previous work like Hist2gene [10] and Hist2ST [11] did not perform well. We found that STimage model with fine-tuning option improves the model performance on small training dataset compared with models using ViT for feature extraction. For a robust comparison, we used the prediction of each of the top 1000 Highly Variable Genes (HVGs), and computed both Pearson correlation and Moran's I scores for each gene. Across 1000 genes, we consistently found that STimage outperformed the other three models. We further assessed the model robustness by testing the prediction performance for each of the 1000 genes and compared it with the current three methods available (Figure 2f). The trained model was on the public company's dataset and Swarbrick's dataset, while the testing was for an independent dataset generated by a lab in Europe, containing data from 36 tissue samples. STimage was benchmarked with three other approaches, including two software STnet and His2gene and one alternative option with ViT_NB. Consistently across all samples tested STimage was the best performer, except for the performance in six (out of 36) samples, where STimage performed as well as the top performers.

Of note, methods implementing ViT (e.g., Hist2gene and Hist2ST), which builds an attention map for spots across the tissue sections, requires more computer memory. It is challenging to run when there are more measurements in tissue, a common scenario enabled by recent development of spatial transcriptomics. STimage model implemented the CNN as an image extractor which treated each measurement as an individual data point, therefore, it can be easily extended to more advanced ST technologies where there are more measurements in each tissue slide. STimage also made use of pre-trained models, which are not applied by ViT.

A classification model can be added top produce complementary information

In addition to predicting continuous gene expression values, STimage also implements a classification model, allowing to predict on the H&E image, of the location with high or low expression of a gene (Figure S4). In this model, the continuous gene expression value was split into two categories, high vs low expression, which is then used for training and evaluation. This approach can be complementary to the model that predicts continuous values as presented in (Figure 1). Comparing the prediction for COX6C in Figure 1c-d with that in Figure S4c-d, it appeared that the classification could be more specific to the image spots that are highly expressed (Figure S4a) but can also be noisier at the spots with the medium expression (Figure S4b).

Interpretability analysis

Moreover, we investigated the interpretability of the model by assessing feature importance using two explainability methods, Linear Interpretability Model-agnostic Explanations (LIME) [23] and Shapley Additive Explainations (SHAP) [24] (refer to Methods), (Figure S1f,S5). We found that the image segments corresponding to nuclei regions contributed more to the prediction (Figure S1f). Further, features that are most predictive for the expression of different genes are not the same tissue segments, for example for COX6C and KRT5 genes, suggesting differential contributions to the model prediction. Using detailed histopathological annotation, we found that regression results had a consistent pattern in the tissue segments that contribute to predicting the expression of cancer (GNAS) or immune genes (C3)(Figure S5). The two genes were selected as they were highly contrasted in spatial expression pattern (Figure S5a,b) This pattern shows that those spots in the cancer regions (e.g., spots 1 and 2 in Figure S5) had more segments supporting the expression of the cancer genes than for immune genes. In contrast, segments with high scores for predicting C3 genes appear more in the immune spot (e.g., spot 3). The classification, on the other hand, appears to be less distinguishable between segments that support the prediction of high expression of cancer or immune genes. our interpretability analysis approach suggests that incorporating histopathological annotation to interpret LIME and SHAP segments is useful. However, this assessment is still qualitative, and more work is required to connect cell segmentation with independent cell type assessment. In particular, we observed for the classification model, the scores for segments do not appear to be associated with cancer or immune spots (Figure S5). We also created an interactive web application for convenient usage and testing of the STimage model and interpretability (Figure S6).

Conclusion

We developed STimage, a comprehensive machine learning software enables the estimation of gene expression for pathological tissue images. STimage regression model integrates deep learning and statistical methods to estimate the distribution of gene expression in ST data and to quantify prediction uncertainty. We benchmarked STimage regression model with existing methods in both leave-one-out experiment and out-of-distribution scenarios. STimage regression model outperforms other methods in most cases. STimage also implemented a classification model which can predict high/low expression of a marker gene or gene list and LIME and SHAP interpretation models to detect meaningful tissue/cell features that contribute to the model prediction. Overall, we believe STimage contributes to making an important progress in applying machine learning for digital pathology.

Declarations

Ethics approval and consent to participate

Not applicable

Availability of data and materials

The datasets analysed in this study are publicly available and can be download from UQ e-space via `https://doi.org/10.48610/4fb74a9`. STimage software and web-app tools are avaliable via GitHub `https://github.com/BiomedicalMachineLearning/STimage`

Acknowledgments

We thank Prof. Alex Swarbrick for kindly providing their spatial transcriptomics data. We thank the staff at the University of Queensland sequencing facility and the School of Biomedical Science imaging facility for their support in spatial transcriptomics sequencing.

Funding

This research was partially supported by the Australian Research Council through an Industrial Transformation Training Centre for Information Resilience (IC200100022). QHN is supported by Australian Research Council (ARC DECRA grant DE190100116), National Health & Medical Research Council (NHMRC Project Grant 2001514), and NHMRC Investigator Grant (GNT2008928).

Competing interests

M.T. and S.M. were employed by Max Kelsen, which is a commercial company with an embedded research team. The remaining authors have declared no competing interest.

Author contributions

Q.N. and X.T. conceived the project. Q.N., X.T., O.M. and S.M. designed the algorithm. S.M. led the uncertainty quantification analysis. X.T. and O.M. wrote the software. K.T. and P.S. performed pathological annotation. F.R., M.T. contributed to the analysis. All authors have contributed to the writing of the manuscript.

Supplementary data

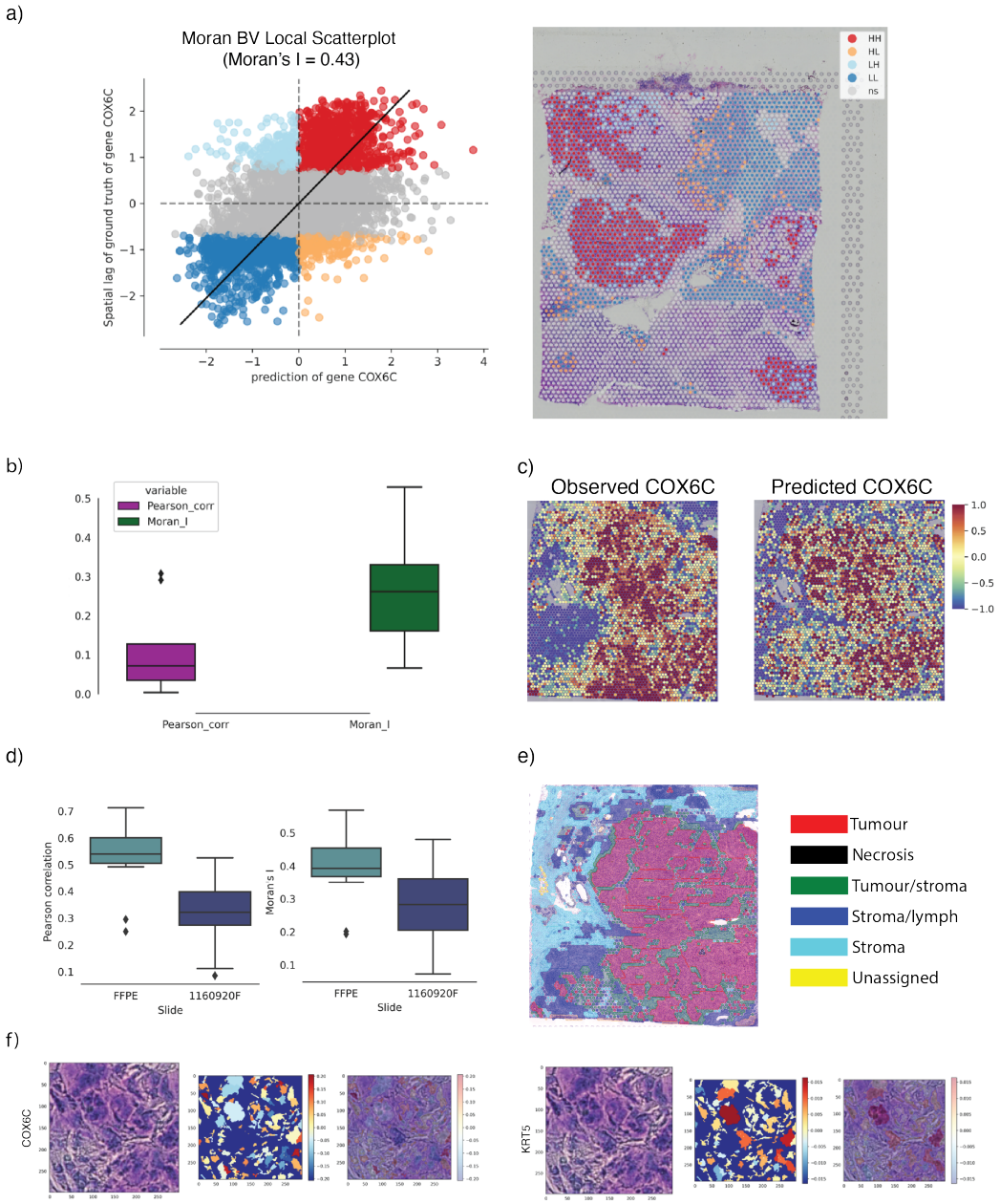

Fig. S1. Performance of STimage CNN-NB regression model. **a**, Spatial autocorrelation using Moran's I index for predicted highly abundant cancer marker COX6C. The points in the scatter plot are also shown in the tissue plot. HH indicate concordance with the high predicted and high observed values. LL = Low-Low, H-L = High-Low, L-H = Low-High, and ns = non-significant. **b**, Pearson correlation and Moran's I were calculated for predicted and observed 14 cancer and immune markers. **c**, Model performance for external test dataset. **d**, Pearson correlation and Moran' I values for predicted and observed 14 cancer and immune markers for the external test dataset. **e**, Ground truth cell type label for external test dataset, where the label was defined by a pathologist as described in the publication [14]. **f**, LIME model interpretation. The model marks different image segments by scores contributing to predicting gene expression, where red colour suggests important pixels for predicting high expression and blue colour as important for low expression

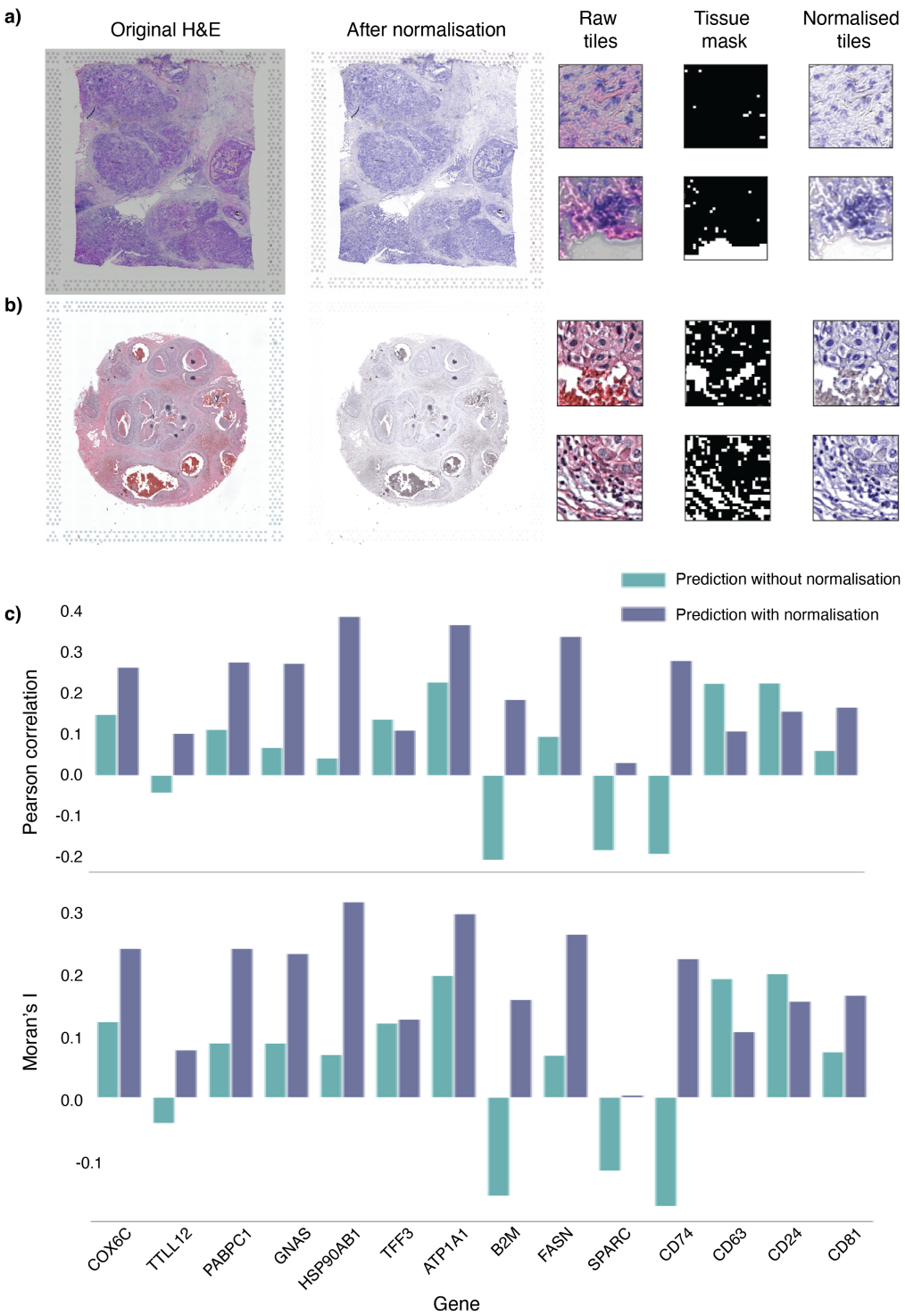

Fig. S2. Preprocessing of the H&E image input before running the STimage model. a) Normalisation and tile removal for fresh frozen tissues. b) Similar to a, but for FFPE tissue. Tissue masking was done by OpenCV2, showing in black areas in the image (white areas indicate regions without tissue covered). A tile with tissue coverage less than 70% is removed. Vahadane [15] stain normalisation is performed (as a default method) to correct for background staining of all images to the same level of a template image. Images before and after normalisation are shown. Comparison of model performance with and without stain normalisation.

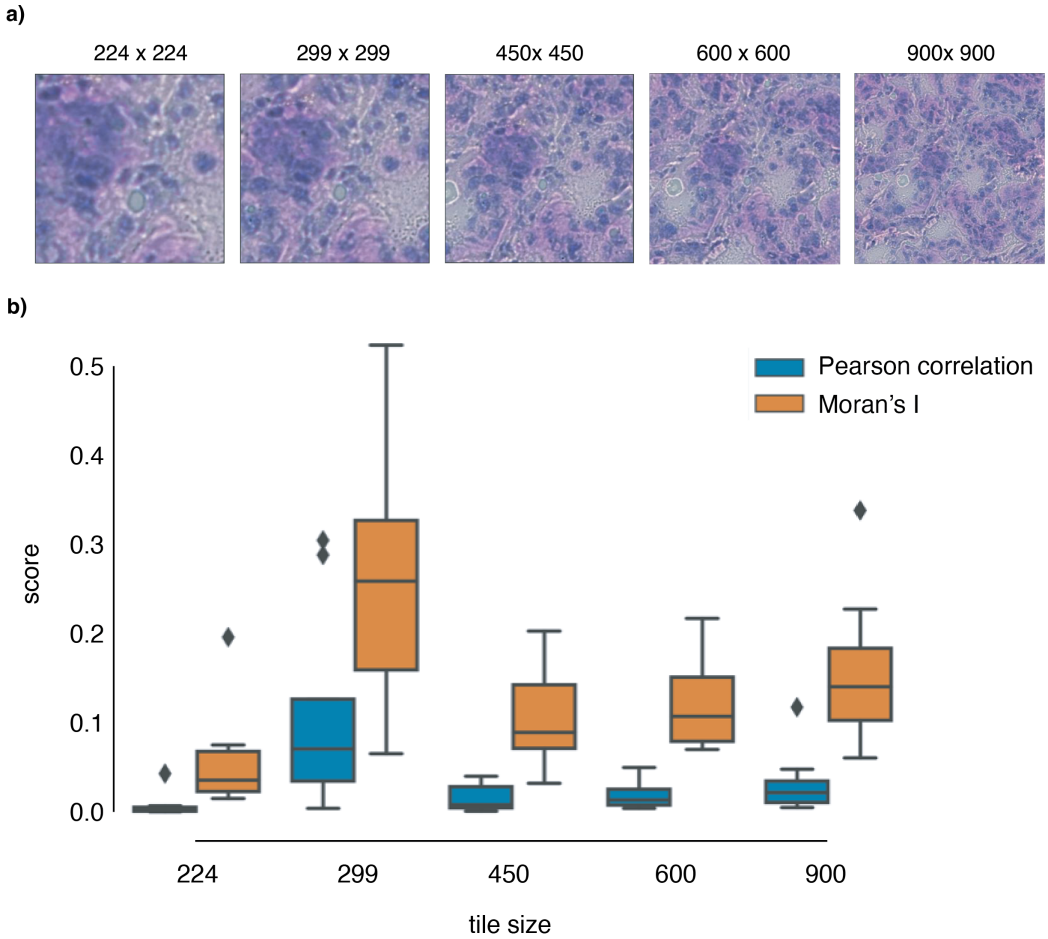

Fig. S3. Optimisation of tile size and its effects for model performance. **a**, options for tile size includes 244x244, 299x299, 450x450, 600x600 and 900x900. **b**, the model performance for different tile sizes was evaluated by Pearson correlation and Moran's I statistics.

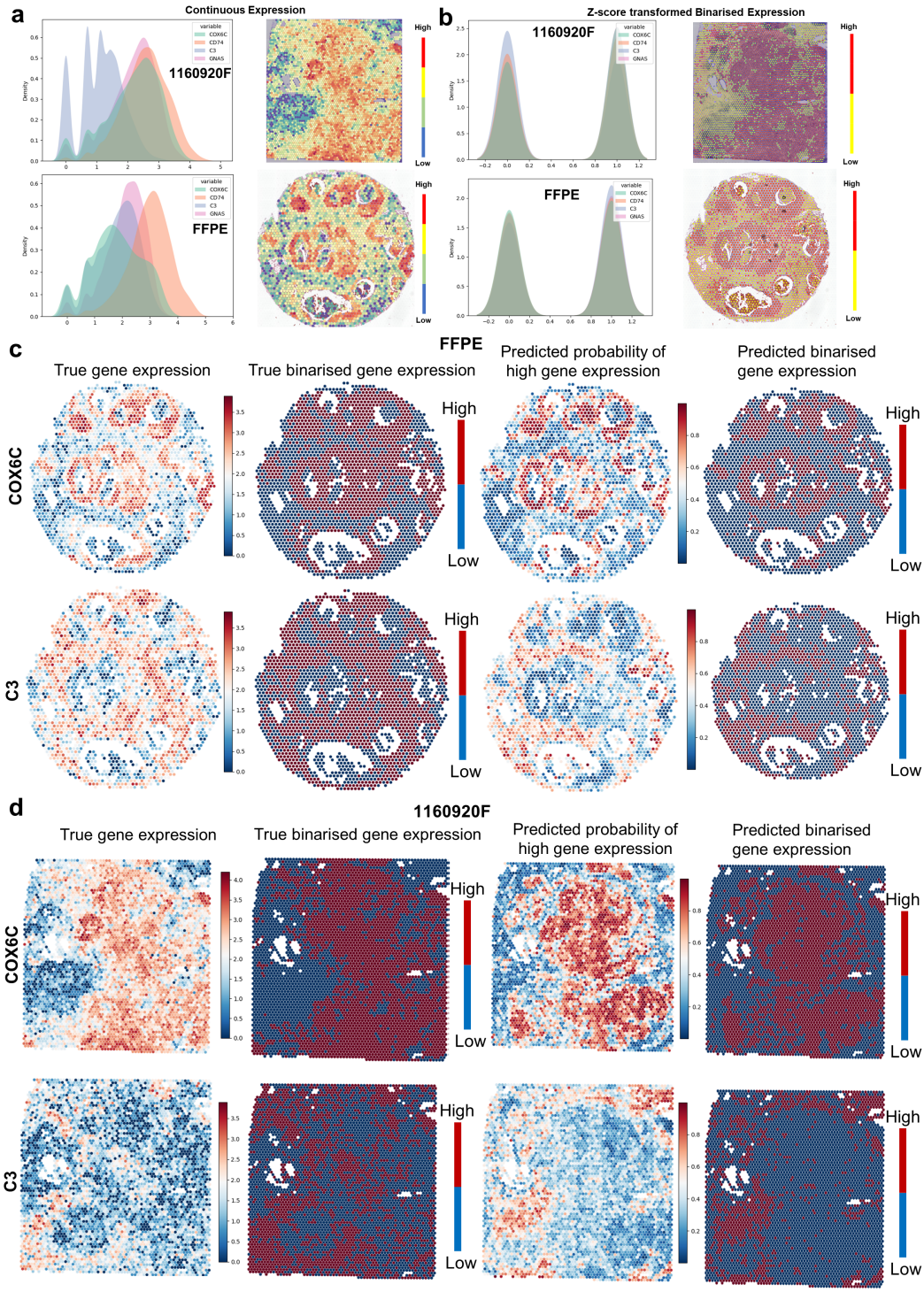

Fig. S4. STimage classification model predicts the probability of a spot with high or low expression. **a,b** Density plot of four genes (COX6C, CD74, C3 and GNAS) and the gene expression plot of COX6C for continuous gene expression for two test samples is shown in **a**. Similarly, the density plot and gene expression plot for binarised gene expression plot is shown in **b**. **c** Prediction for a FFPE dataset. From left to right, experimental data (true gene expression), binarised data (classification of high vs. low spots; red and able as high and low expression categories, respectively), class probability from softmax layer in the neural network, and the predicted binary classes. Two genes are shown, COX6C as a cancer marker and C3 as an immune marker. **d** Similar to c, but for a spatial dataset prepared by a different protocol (for a fresh frozen sample)

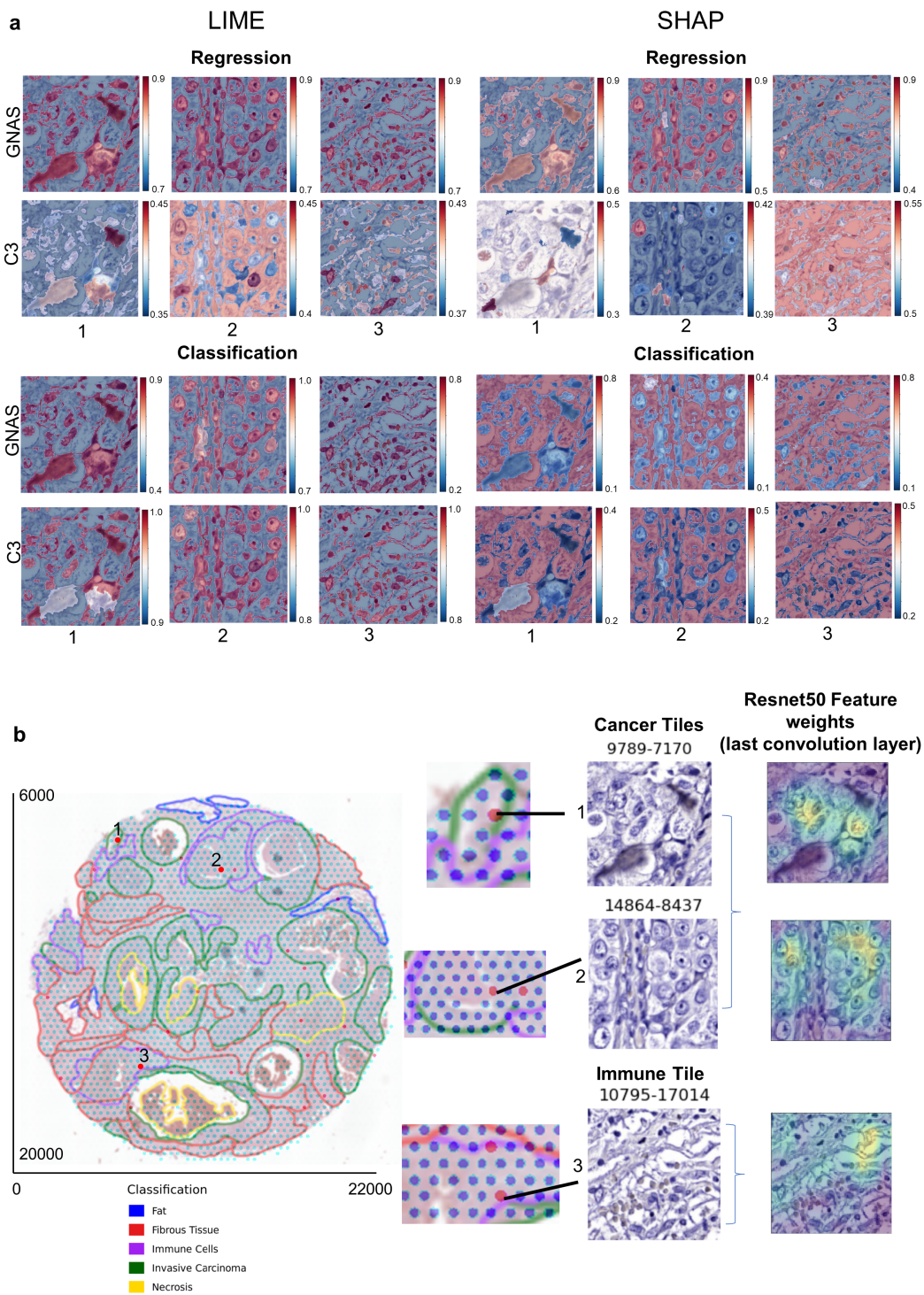

Fig. S5. Model Interpretability for STimage. a) Results from two methods, LIME and SHAP for regression and classification models are shown. The models were trained on a pair of cancer (GNAS) and immune (C3) gene with low spatial auto-correlation i.e., low Moran's Index. The red segments are important features for predicting the gene expression. The LIME and SHAP were run for 25 random tiles and then LIME and SHAP scores were normalised for these 25 tiles. b) Displaying the locations of tiles in with pathologist annotation on H&E and the ResNet50 feature weight heatmaps from the last convolutional layer overlayed on corresponding image tiles.

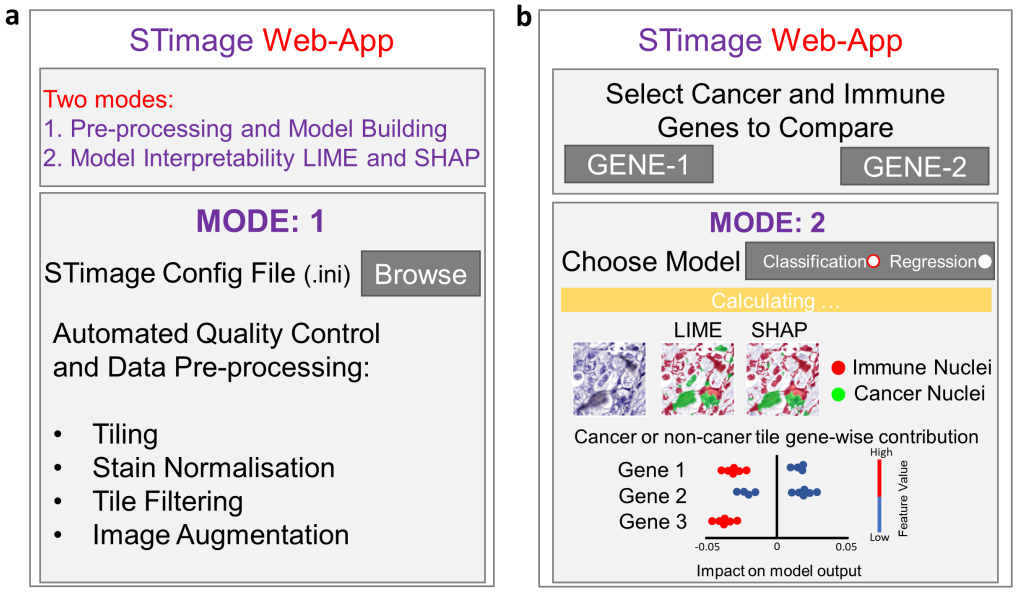

Fig. S6. STimage Web Application. a) The first mode of the web app converts raw Visium data to anndata, performs quality control steps. These steps involve filtering genes and image tiles, performing stain normalisation and allowing the user to build their own machine-learning model by submitting a config file (an example config file is available on our github). b) In the second mode, users can visualise important features of an image tile that are responsible for the prediction of gene expression. It also allows the users to check the contribution of each gene for classifying an image tile as cancer or non-cancer. The source code is available at

References

1. Joann G Elmore, Raymond L Barnhill, David E Elder, Gary M Longton, Margaret S Pepe, Lisa M Reisch, Patricia A Carney, Linda J Titus, Heidi D Nelson, Tracy Onega, Anna N A Tosteson, Martin A Weinstock, Stevan R Knezevich, and Michael W Piepkorn. Pathologists' diagnosis of invasive melanoma and melanocytic proliferations: observer accuracy and reproducibility study. *BMJ*, 357:j2813, June 2017.
2. Anoop P Patel, Itay Tirosh, John J Trombetta, Alex K Shalek, Shawn M Gillespie, Hiroaki Wakimoto, Daniel P Cahill, Brian V Nahed, William T Curry, Robert L Martuza, David N Louis, Orit Rozenblatt-Rosen, Mario L Suvà, Aviv Regev, and Bradley E Bernstein. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science*, 344(6190):1396–1401, June 2014.
3. Andrew J Evans, Thomas W Bauer, Marilyn M Bui, Toby C Cornish, Helena Duncan, Eric F Glassy, Jason Hipp, Robert S McGee, Doug Murphy, Charles Myers, Dennis G O'Neill, Anil V Parwani, B Alan Rampy, Mohamed E Salama, and Liron Pantanowitz. US food and drug administration approval of whole slide imaging for primary diagnosis: A key milestone is reached and new questions are raised. *Arch. Pathol. Lab. Med.*, 142(11):1383–1387, November 2018.
4. Joel Saltz, Rajarsi Gupta, Le Hou, Tahsin Kurc, Pankaj Singh, Vu Nguyen, Dimitris Samaras, Kenneth R Shroyer, Tianhao Zhao, Rebecca Batiste, John Van Arnam, Cancer Genome Atlas Research Network, Ilya Shmulevich, Arvind U K Rao, Alexander J Lazar, Ashish Sharma, and Vésteinn Thorsson. Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images. *Cell Rep.*, 23(1):181–193.e7, April 2018.
5. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, February 2017.
6. M Janda and H P Soyer. Can clinical decision making be enhanced by artificial intelligence? *Br. J. Dermatol.*, 180(2):247–248, February 2019.
7. Benoît Schmauch, Alberto Romagnoni, Elodie Pronier, Charlie Saillard, Pascale Maillé, Julien Calderaro, Aurélie Kamoun, Meriem Sefta, Sylvain Toldo, Mikhail Zaslavskiy, Thomas Clozel, Matahi Moarii, Pierre Courtiol, and Gilles Wainrib. A deep learning model to predict RNA-Seq expression of tumours from whole slide images. *Nat. Commun.*, 11(1):3877, August 2020.
8. Patrik L Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson, Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353(6294):78–82, July 2016.
9. Bryan He, Ludvig Bergenstråhle, Linnea Stenbeck, Abubakar Abid, Alma Andersson, Åke Borg, Jonas Maaskola, Joakim Lundeberg, and James Zou. Integrating spatial gene expression and breast tumour morphology via deep learning. *Nature biomedical engineering*, 4(8):827–834, 2020.
10. Minxing Pang, Kenong Su, and Mingyao Li. Leveraging information in spatial transcriptomics to predict super-resolution gene expression from histology images in tumors. *bioRxiv*, 2021.
11. Yuansong Zeng, Zhuoyi Wei, Weijiang Yu, Rui Yin, Yuchen Yuan, Bingling Li, Zhonghui Tang, Yutong Lu, and Yuedong Yang. Spatial transcriptomics prediction from histology jointly through transformer and graph neural networks. *Brief. Bioinform.*, 23(5), September 2022.

12. Taku Monjo, Masaru Koido, Satoi Nagasawa, Yutaka Suzuki, and Yoichiro Kamatani. Efficient prediction of a spatial transcriptomics profile better characterizes breast cancer tissue sections without costly experimentation. *Sci. Rep.*, 12(1):4133, March 2022.
13. Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning, 2017.
14. Sunny Z Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I Weisenfeld, Cedric R Uytingco, Jennifer G Chew, Zachary W Bent, Chia-Ling Chan, Vikkitharan Gnanasambandapillai, Charles-Antoine Dutertre, Laurence Gluch, Mun N Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E Powell, Stephen R Williams, X Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M Perou, and Alexander Swarbrick. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.*, 53(9):1334–1347, September 2021.
15. Abhishek Vahadane, Tingying Peng, Amit Sethi, Shadi Albarqouni, Lichao Wang, Maximilian Baust, Katja Steiger, Anna Melissa Schlitter, Irene Esposito, and Nassir Navab. Structure-preserving color normalization and sparse stain separation for histological images. *IEEE Trans. Med. Imaging*, 35(8):1962–1971, August 2016.
16. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.
17. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320, 2005.
18. Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. *Advances in neural information processing systems*, 27, 2014.
19. Serge Beucher. The watershed transformation applied to image segmentation. *Scanning Microscopy*, 1992(6):28, 1992.
20. Ludvig Bergenstråhle, Bryan He, Joseph Bergenstråhle, Xesús Abalo, Reza Mirzazadeh, Kim Thrane, Andrew L Ji, Alma Andersson, Ludvig Larsson, Nathalie Stakenborg, Guy Boeckxstaens, Paul Khavari, James Zou, Joakim Lundeberg, and Jonas Maaskola. Super-resolved spatial transcriptomics by deep data fusion. *Nat. Biotechnol.*, 40(4):476–479, April 2022.
21. Samual MacDonald, Helena Foley, Melvyn Yap, Rebecca L Johnston, Kaiah Steven, Lambros T Koufariotis, Sowmya Sharma, Scott Wood, Venkateswar Addala, John Pearson, et al. Generalising uncertainty improves accuracy and safety of deep learning analytics applied to oncology. *bioRxiv*, pages 2022–07, 2022.
22. Samual MacDonald, Kaiah Steven, and Maciej Trzaskowski. Interpretable ai in healthcare: Enhancing fairness, safety, and trust. In *Artificial Intelligence in Medicine: Applications, Limitations and Future Directions*, pages 241–258. Springer, 2022.
23. Marco Túlio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should I trust you?": Explaining the predictions of any classifier. *CoRR*, abs/1602.04938, 2016.
24. Scott M Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30, 2017.