An investigation of attention mechanisms in histopathology whole-slide-image analysis for regression objectives

Philippe Weitz Yinxi Wang Johan Hartman Mattias Rantalainen
Karolinska Institutet
{philippe.weitz, yinxi.wang, johan.hartman, mattias.rantalainen}@ki.se

Abstract

Analysis of whole-slide-images (WSIs) of histopathology tissue sections remains challenging due to the gigapixel scale of these images, which often necessitates their division into smaller image tiles. Recently, attention mechanisms have been successfully applied to alleviate the tile-to-slide challenges for classification tasks based on WSIs. In this study, we explore the potential of attention mechanisms in regression settings, by comparing four modelling approaches, two of which use attention mechanisms. We evaluate these models both in a simulated experiment using the MNIST data set, and in real histopathology data sets focused on prediction of gene expression levels from WSIs, including an analysis of the local prediction performance using spatial transcriptomics. The MNIST simulation demonstrates that if only a small proportion of instances in a set of images contribute to the set-level regression label, attention mechanisms may be preferable to commonly applied weakly supervised models. When predicting gene expression from WSIs, the differences in performance between the models that we investigated were small. Nevertheless, we found some evidence that attention mechanisms may be more sensitive to domain shifts. In the regression-based task of gene expression prediction, the prediction performance in the present study appears to be limited by other factors rather than by the choice of modelling approach. Nevertheless, attention mechanisms appear promising for regression objectives and warrant further investigation.

1. Introduction

The emergence and application of deep learning models such as convolutional neural networks (CNNs) as well as the increasing availability of digital whole-slide-images (WSIs) of histopathological tissue sections has lead to remarkable advances in computational pathology in recent time. However, the application of deep learning models in computational pathology at scale remains challenging, since current computer hardware and especially graphical processing units (GPUs) are not yet equipped to operate on entire WSIs due to their gigapixel scale. To circumvent these challenges, WSIs are typically divided into smaller image patches(tiles) to fit into GPU memory. For some labels, such as pixel-level annotations, e.g. the location and type of cells, or semantic segmentations, e.g. the distinction between benign and cancer regions, this division into image patches does not affect the relationship between image and label. However, in cases with slide-level labels only, e.g. histological grade, treatment response or survival time, the relationship between tile level and slide level labels is more complex.

A common naïve solution to this is to assign the slide-level label to all image tiles, which is often referred to as weakly supervised learning in computational pathology. Despite the surprisingly good performance of this approach in many studies, it is desirable to find a solution that better models the relationship between WSI-level label and the individual tiles. For example, if a large proportion of tiles does not contribute to the WSI-level label, this method may result in poor prediction performance, since the non-contributing tiles essentially add noise to the WSI-level predictions.

Different methods have been proposed to alleviate this by aiming to directly predict on the WSI. Tellez *et al*. [19] suggested to compress WSIs by extracting a feature representation for each tile using a pretrained CNN and subsequently training a secondary CNN on an image that is composed of these feature vectors, such that each pixel has one channel for each extracted feature. More recently, Pinckaers *et al*. [15] introduced streaming neural networks (NNs) to histopathology image analysis. Both of these methods have the advantage of preserving spatial information. For classification tasks, it has been proposed to formulate the relationship between tile-level and slide-level label as a multiple instance learning (MIL) problem, where for binary classification, a set or bag of instances is positive if the bag contains at least one positive instance. Campanella *et al*. [4] demonstrated that a hybrid of MIL and weakly super-

vised learning, in which the top K tiles of a WSI are considered, can yield promising results. However, this method appears to require an exceptionally large data set. Ilse *et al*. [8] proposed an attention mechanism for end-to-end MIL model optimization and applied it in the histopathology setting. This attention mechanism generates a slide-level representation, which is then classified by a fully connected neural network. Lu *et al*. [11, 12] recently expanded this approach to multi-class classification and demonstrated that the attention attribution corresponds well to the relevant image regions if these are known, e.g. in metastases detection in lymph nodes. Although these attention-based models do not account for spatial organisation of tiles or the interdependence of different image regions, they nevertheless have increased prediction performance and data efficiency.

While there is a natural interpretation between bag labels and instance labels for many classification tasks in histopathology that is well described by the MIL approach, regression tasks cannot be posed analogously. However, the attention mechanisms that bridge tile- and slide-level labels may also prove useful in this setting if different tiles have different contributions to the slide-level label. While some regression tasks use whole-tumor-averages as the slide-level label, i.e. the percentage of Ki67-positive cells for Ki67 scoring, other labels are based on hotspots, such as the mitotic counts in breast cancer. For other regression tasks, such as gene expression prediction, the relationship between tile- and slide-level labels is currently unknown, although several recent studies have demonstrated that gene expression can be predicted from WSIs by using bulk RNA-sequencing expression estimates as a weak label for image tiles [6, 18, 21, 22]. The nature of intra-tumor heterogeneity of gene expression is currently an active area of research, and it is at this point in time unknown if the bulk average expression profile that is measured by RNA-seq is driven by hotspots of high expression, or driven by global tumour expression changes. Different types of variability could mean that different modelling approaches are effective in different scenarios and for different genes, particularly considering the large dynamic range of gene expression values. Challenges for model-based predictions of expression could occur if there are substantial saturation effects with respect to morphology changes in high expression hotspots, or for transcripts that only have a weak association between morphology and expression levels, even if the expression is relatively homogeneous across the WSI.

Even though adaptively weighting tiles appears promising for regression objectives as well because of the potential of different contributions of tiles to the slide-level label, to the best of our knowledge, attention mechanisms have not yet been applied to regression objectives in the histopathology domain. In this study, we therefore investigate four different modelling approaches that assume different relationships between tile-level labels and slide-level label, two of which use attention mechanisms. We evaluate these models both with a simulated example using MNIST data, as well as through predicting gene expression levels from WSIs of H&E stained breast cancer tumor sections. We validate our findings using both an independent cohort of patients as external test data, as well as with spatial transcriptomics.

2. Materials & methods

In this study, we compare four different modelling approaches that differ in their assumptions regarding the relationship between individual tile-level labels and slide-level labels. We compare these models both with a simulated MNIST experiment, as well as through predicting gene expression levels from H&E stained WSIs.

2.1. Models

All four models compared here share the same fully connected NN structure that predicts a regression output. The first two layers of this structure are two fully connected layers that each reduce the number of input features by half. Model predictions are generated by a final fully connected layer that expects inputs with the size of a quarter of the initial input to the fully connected NN. The input of this final layer differs between the four investigated models.

The two models with an attention mechanisms use the same attention network structure as described in [8, 11, 12] for models with a single output. The purpose of the attention network is to adaptively obtain an attention weight for each instance in a set of images or their corresponding feature representations. This set of feature representations is denoted by $h \in \mathbb{R}^{n \times k}$. This number is held constant during training (see Table 1) to allow for batches encompassing multiple sets, which for WSIs is accomplished through oversampling if a WSI has fewer than k tiles. During prediction, k corresponds to the total number of tiles in the respective WSI. The attention network itself consists of two fully connected layers with parameters $U, V \in \mathbb{R}^{\frac{n}{2} \times n}$ and an independent set of weights $W \in \mathbb{R}^{1 \times \frac{n}{2}}$. Here, n corresponds to the number of input features to the attention network. With $\odot$ as the element-wise product, the vector of attention weights $a \in \mathbb{R}^k$ is

$$a = \frac{\exp(W \tanh(Vh) \odot \text{sigmoid}(Uh)}{\sum_{j=1}^{k} \exp(W \tanh(Vh_j) \odot \text{sigmoid}(Uh_j)}, \quad (1)$$

where $h_j \in \mathbb{R}^{n \times 1}$ is the feature representation of the j-th image patch. The exponential functions and sum in Equation 1 correspond to the *softmax* function. Multiplying either a vector of predictions or a matrix of features representations with a therefore results in an attention-weighted set-wide average. With this attention mechanism, we compared the following four model structures:

1. Attention-weighted average of features (AF). As proposed in [8, 12], a slide-level feature representation is generated by obtaining an attention-weighted average of tile-level feature representations. A fully connected NN is then trained to predict the slide-level label from this slide-level aggregate.

2. Attention-weighted average of predictions (AP). While the AF model generates a slide-level representation by aggregating feature vectors, this model predicts the slide-level label from each tile and uses an attention-weighted average of individual tile-level predictions to generate the slide-level prediction.

3. Mean of features (MF). This model uses the mean of all feature vectors as a slide-level representation, which is used to train a NN to predict slide-level labels. This model is equivalent to the AF model if all attention weights of the AF model were equivalent.

4. Mean of predictions (MP). This approach corresponds to the naïve, weakly supervised approach in which the slide-level label is assigned to each tile. Slide-level predictions are the mean of all tile-level predictions. This approach is equivalent to the AP model if all attention weights of the AP model were equivalent.

Models that use the mean of either all instance-level features (MF) or predictions (MP) assume that all instances contribute equivalently to the set-level label, whereas the models that use an attention mechanism (AF, AP) allow for different contributions of instances. Models that aggregate features (AF, MF) generate a set-level feature representation, whereas models that aggregate predictions (AP, MP) generate predictions based on instance-level features and aggregate instance-level predictions to set-level predictions. We obtained instance-level predictions for the AF, MF models through regarding each image as a set of a single instance.

Between the MNIST simulation and the histopathology application, the networks differ in the number of neurons per layer, depending on the number of CNN-extracted features. The MNIST models furthermore have a CNN feature extractor network that is optimized end-to-end alongside the fully connected NN components. The structure of the MNIST feature extractor corresponds to examples on the PyTorch website [16] and yields 320 features. To reduce the computational cost, the histopathology models use 512 features that were extracted with a ResNet18 model [7] with ImageNet [17] weights and are not trained end-to-end.

2.2. MNIST simulation

In order to investigate the four models in a controlled experiment, we devised a simulation based on the MNIST data set [10]. The purpose of this simulation is to compare how well the four described models are able to learn instance and set-level labels from set-level labels in a regression setting. We investigate this under varying fractions of instances per set that contribute to the set label. To this end, MNIST images were randomly grouped into sets of images whose labels are defined as the mean of all individual MNIST image labels $\in [0, 9]$ in that set. We then randomly generated noise images whose pixel values were randomly drawn from a uniform distribution $U(0, 255)$ and added these to the image sets, such that each set size is 32, which we chose to obtain a reasonably large number of sets. These noise images do not contribute to the set label. We compared four different proportions of noise images per set, 0, 0.25, 0.5 and 0.75. Using MNIST images and noise to generate image sets has the advantage that the true label of each contributing instance is known. It is therefore possible to not only compare the predictions of the four models on a set level, but also on an instance level.

For model training, the MNIST training data (60,000 images) was split into a training set (48,000 images) and a validation set (12,000 images). The validation set was used for parameter tuning and early stopping based on the validation loss. Models were trained using the *Adam* optimizer for up to 100 epochs with an early stopping patience of 10 epochs, a learning rate of 0.0001 and a batch size of one set, using the *mean-absolute-error* (MAE) as the cost function. Model performance was then assessed on analogously generated sets of images based on the MNIST test set (10,000 images). We repeated this experiment 100 times per model and proportion of noise, iterating over 100 random seeds per configuration. The random seeds determine the split into training and validation data and the generation of image sets and noise images. The CNN feature extractor and model parameters were initialized with the same random seed for each of the experiment configurations and repetitions where applicable to reduce random effects and allow pair-wise comparisons between models with the same seed and configuration. To make the analysis and interpretation of results comparable between the MNIST simulation and the WSI application, we used Spearman correlation as the primary performance metric to evaluate the set and instance level prediction performances of the models at the different noise proportions.

2.3. Histopathology application

In the histopathology domain we investigate prediction of gene expression, which is a regression problem, and evaluate predication performance of the four modelling approaches on both slide-level and on instance (tile) level, with local prediction performance evaluated by spatial transcriptomics.

2.3.1 Data

This study includes WSIs from four different cohorts of female breast cancer patients, two of which were exclusively used for model validation. The first cohort, Clinseq [20], consists of 270 patients. We furthermore selected 721 patients from the publicly available TCGA BRCA [5] data set based on the availability of additional clinical information. We randomly selected 697 patients from Clinseq and TCGA for model training, 122 for validation and 172 as an internal test set. As an external test set, we used 350 patients from the ABiM study [3]. The fourth data set consists of 22 patients for which spatial transcriptomics data is available. For all patients from Clinseq, TCGA and ABiM, bulk RNA-sequencing estimates of gene expression values for at least the majority of currently known protein coding genes are available. TCGA and Clinseq data was pre-processed using an identical protocol. To reduce remaining batch effects, each of these datasets were also median centered (gene-wise). We randomly selected 125 transcripts (100 for validation, 25 for hyperparameter tuning) out of 1011 genes that have previously been shown [21] to be predictable from WSIs of H&E stained breast cancer sections for analysis in this study.

In order to investigate whether any of the modelling approaches is superior in learning local information based on slide-level labels, we also evaluated the models with a data set that has local gene expression values from spatial transcriptomics (ST) available. This ST data set consists of 22 WSIs of H&E stained tissue sections of breast tumors. For each of these sections, a consecutive tissue section was used for spatial transcriptomics analysis of 84 expression levels with the GeoMx DSP platform and the NanoString nCounter® instrument (GeoMx Immune Pathways Panel, NanoString Technologies, Seattle, WA). For each tissue section, 12 ROIs of 600μm×600μm were expression profiled (264 ROIs in total). The ST tissue sections were stained with fluorescent stains targeting PANCK, SMA, CD45 and DNA and manually registered to the WSIs of the H&E stained sections. Out of the 84 expression values per ROI that were obtained, 6 were used as negative controls. The remaining 78 gene expressions were normalized based on the average expression value of these 6 transcripts. The remaining transcripts were then $\log_2$-transformed. Two transcripts were subsequently excluded due to low variance (< 0.001), and two expression features were excluded as they correspond to the average expression of several genes. Further 74 genes were included in our analysis, resulting in 174 for analysis and 25 for hyperparameter tuning.

Hyperparameter	**Evaluated**
CNN feature extractor	InceptionV3, **ResNet18**
Learning rate	**1e-3**, 1e-4
Training set size k	100, 250, 500, 1000, **2000**
Number of sets per batch	**1**, 2, 4, 8, 16, 32, 64

Table 1. Hyperparameters that were evaluated on a set of 25 randomly selected transcripts, with selected hyperparameters in bold. The 25 transcripts used to identify the optimal hyperparameters were not part of the further analysis. Optimal hyperparameters did not vary for any of the four models.

2.3.2 Image preprocessing

For each WSI, tissue regions were detected by applying Otsu thresholding [14] to the HSV saturation channel and a threshold of 0.75 to the hue channel. All WSIs were tiled into image patches of 598×598 pixels (271μm×271μm) at 20X resolution. Tiles with a variance of less than 500 after Laplacian filtering were assumed to be out of focus and excluded. All image tiles were then normalized with the method described by Macenko *et al*. [13]. Subsequently, regions of invasive cancer were detected with a cancer detection CNN. Further details of the WSI scanners used, the image preprocessing and the cancer detection model can be found in [21], which uses the same WSI data sets as this study. For each tile that belongs to a predicted cancer region, we extracted a feature representation that comprises 512 features with a ResNet18 model with ImageNet weights.

2.3.3 Model optimization

Models were optimized with tiles from WSIs of the 697 patients in the training set, again using *Adam* as the optimizer and MAE as the loss function. The validation set was used for hyperparameter tuning and early stopping based on the validation loss, with an early stopping patience of 10 epochs and a maximum of 100 epochs. Hyperparameters were tuned with a randomly selected set of 25 transcripts that was not included into the further analysis. The explored and selected hyperparameters are provided in Table 1. For each of the 174 transcripts that we planned to evaluate in the test data, we fitted each of the four models, resulting in 696 fitted models.

2.3.4 Model validation

Slide-level predictions were validated with an internal test set of 172 patients consisting of patients from Clinseq and TCGA, as well as the external ABiM data set. Since the gene expression data of the ABiM cohort was preprocessed differently than Clinseq and TCGA, we used Spearman correlations between slide-level predictions and bulk gene ex-

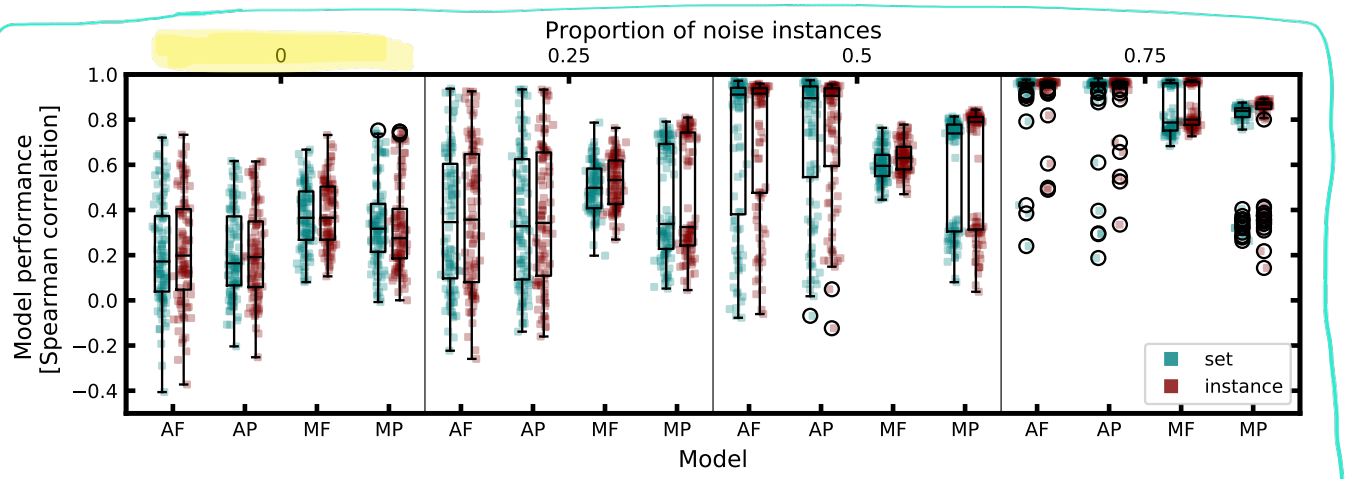

Figure 1. MNIST simulation results with boxplots of the distributions of Spearman correlations over 100 random seeds for each model and proportion of noise. The Spearman correlation for each random seed is overlaid as a blue rectangle for the set-level evaluation and in red for the instance-level. Boxes indicate interquartile ranges with the median marked as a horizontal line inside the respective box. Whiskers indicate 1.5 interquartile ranges, circles denote outliers. Table 2 provides the corresponding FDR-adjusted p-values for pairwise comparisons with Wilcoxon tests.

	0% noise instances				25% noise instances				50% noise instances				75% noise instances			
	AF	AP	MF	MP	AF	AP	MF	MP	AF	AP	MF	MP	AF	AP	MF	MP
AF-s		0.988	1	1		0.986	1	1		0.999	0.077	0.018		0.503	<0.01	<0.01
AP-s	0.847		1	1	0.85		1	1	0.373		0.011	0.011	1		<0.01	<0.01
MF-s	<0.01	<0.01		0.172	<0.01	<0.01		0.015	0.999	0.999		0.257	1	1		<0.01
MP-s	<0.01	<0.01	1		0.032	0.07	1		0.999	0.999	0.999		1	1	1	
AF-i		0.895	1	1		0.918	1	1		0.999	0.026	0.01		0.67	<0.01	<0.01
AP-i	0.895		1	1	0.928		1	1	0.505		<0.01	<0.01	1		<0.01	<0.01
MF-i	<0.01	<0.01		<0.01	<0.01	<0.01		0.014	0.999	0.999		0.143	1	1		<0.01
MP-i	<0.01	<0.01	1		0.032	0.062	1		0.999	0.999	0.999		1	1	1	

Table 2. MNIST simulation results. FDR-adjusted p-values from one-sided Wilcoxon signed-rank tests for Figure 1. Set-level comparisons are marked with -s, instance-level comparisons with -i. The Wilcoxon tests evaluate whether the null hypothesis that the distribution of models marked with -i or -s is not larger than the distribution of the models without specifier can be rejected.

pression estimates as the primary performance metric to avoid sensitivity to offsets and scaling. All Spearman correlation associated p-values were adjusted for multiple testing using the method described by Benjamini and Hochberg (BH) [2].

The local prediction performance in the ST data set was assessed with linear mixed effect (LME) models, in order to account for variations between individual slides. A single prediction for each of the 12 ST ROIs of each WSI was obtained through considering all tiles of a ROI as a set of instances. One LME model was fitted for each prediction model and transcript, with the model predictions as the fixed effect, the WSI ID as the random effect and the $\log_2$-transformed expression value as the response. In order to obtain a metric that is comparable across LME model fits, we computed the proportion of variance explained by the fixed effect as the primary performance metric for the ST analysis. LME models were fitted with the *lme4* [1] *R* package, respective proportions of variance explained were computed with the *r2glmm R* package using the method described by Johnson [9]. As a secondary metric, we again computed Spearman correlations between model predictions and expression values. In this analysis, we computed one Spearman correlation for each transcript, model and WSI in the ST data set as the correlation between the predictions for the 12 ROIs and the corresponding expression values, which results in a distribution of Spearman correlations for each transcript and model. We only evaluated transcripts in the ST data for which at least one of the models had a Spearman correlation with an FDR-adjusted p-value below 0.01 in both the internal test set and the ABiM data set.

	Internal test set						External test set					
	Spearman correlation		Wilcoxon p-value				Spearman correlation		Wilcoxon p-value			
	Median	p <0.01	AF-2	AP-2	MF-2	MP-2	Median	p <0.01	AF-2	AP-2	MF-2	MP-2
AF-1	0.38	152		1	<0.01	0.909	0.351	153		0.042	< 0.01	0.906
AP-1	0.387	147	0.132		<0.01	0.862	0.35	152	1		0.019	1
MF-1	0.359	151	1	1		1	0.347	152	1	1		1
MP-1	0.369	147	0.862	0.909	<0.01		0.357	156	0.906	0.751	0.042	

Table 3. Results for the internal and external test set. Statistics and p-values for the four models and 174 evaluated transcripts in the internal test set and the external ABiM data set. For each data set, the median and number of transcripts with an FDR-adjusted Spearman correlation associated p-value below 0.01 per model are provided, along the FDR-adjusted p-values from one-sided Wilcoxon signed-rank tests that evaluate whether the median Spearman correlation of the models along the rows, denoted with 1, is larger than of the models listed in the columns, denoted with 2.

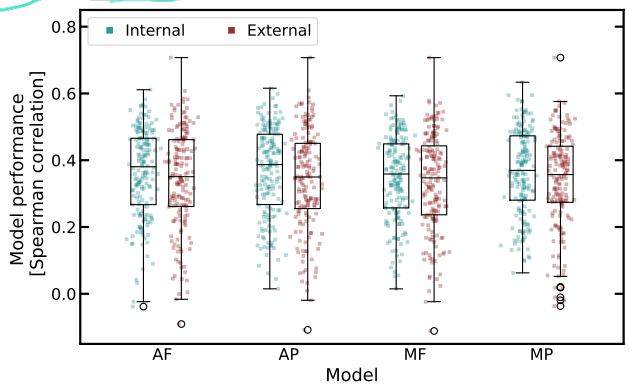

Figure 2. Model performances for the internal and external test set as distributions of Spearman correlations between predicted expression values and RNA-seq for 174 transcripts. Boxplots with the distribution of Spearman correlations for the internal test set are overlaid in blue, red indicates the external test set. Boxplots were generated analogously to those in Figure 1. Corresponding statistics are provided in Table 3.

3. Results

3.1. MNIST simulation

The the model performances (Spearman correlations) on a set and instance-level for the 100 random seeds per model and proportion of noise are depicted in Figure 1. The median Spearman correlation of all models increases with the proportion of noise, or put differently, a lower number of instances that contributes to the set-level label. Within each proportion of noise, we performed one-sided Wilcoxon signed-rank tests between all possible pairs of distributions both for the set level and the instance level. The Wilcoxon tests test the null hypothesis that the distribution of the pairwise differences paired on the random seeds is not lower than zero. The FDR-adjusted p-values from the Wilcoxon tests in Table 2 indicate that the mean models MF and MP outperform their attention counterparts both with regards to the set-level as well as the instance-level predictions. For proportions of noise ≥ 0.5, this is reversed and the models with attention mechanisms outperform the mean models. For the highest proportion of noise, the Wilcoxon tests indicate that the MP model may result in lower model performance than the other three models.

3.2. WSI-level gene expression predictions

We evaluated the association between predicted and RNA-seq estimated expression of 174 transcripts in the internal and the external test sets (Figure 2, Table 3) of the four models, analogously to the MNIST simulation. The null hypothesis of the Wilcoxon tests can be rejected with an FDR-adjusted p-value <0.05 for all models when comparing their distributions to the distribution of the MF model both in the internal test set as well as the external ABiM test set. Otherwise, the Wilcoxon tests give no apparent indication of a difference between the model performances. The median of all models is marginally higher in the internal test data compared to the respective median in the external test set, with a drop in median of 0.029 for the AF model, 0.037 for the AP model, 0.012 for the MF model and 0.012 for the MP model. Performing one-sided Wilcoxon signed-rank tests between the distribution of Spearman correlations of each model in the internal test set and the external test set results in FDR-adjusted p-values of <0.01 for the AF model, <0.01 for AP, 0.019 for MF and <0.01 for MP.

3.3. Spatial expression predictions

Spatial predictions were significant for 42 transcripts for the AF model, 42 for AP, 44 for MF and 39 for MP (FDR-adjusted LME coefficient p< 0.01). Figure 3a) shows the proportions of variance explained by the fixed effect in the LME models for the ten transcripts with the highest proportion of variance predicted among the four compared models. Figure 3b) depicts the corresponding distributions of Spearman correlations. The transcript with the highest proportion of variance predicted for all models corresponds to the immune-related gene MS4A1, which encodes a B-lymphocyte surface molecule. The difference in proportion

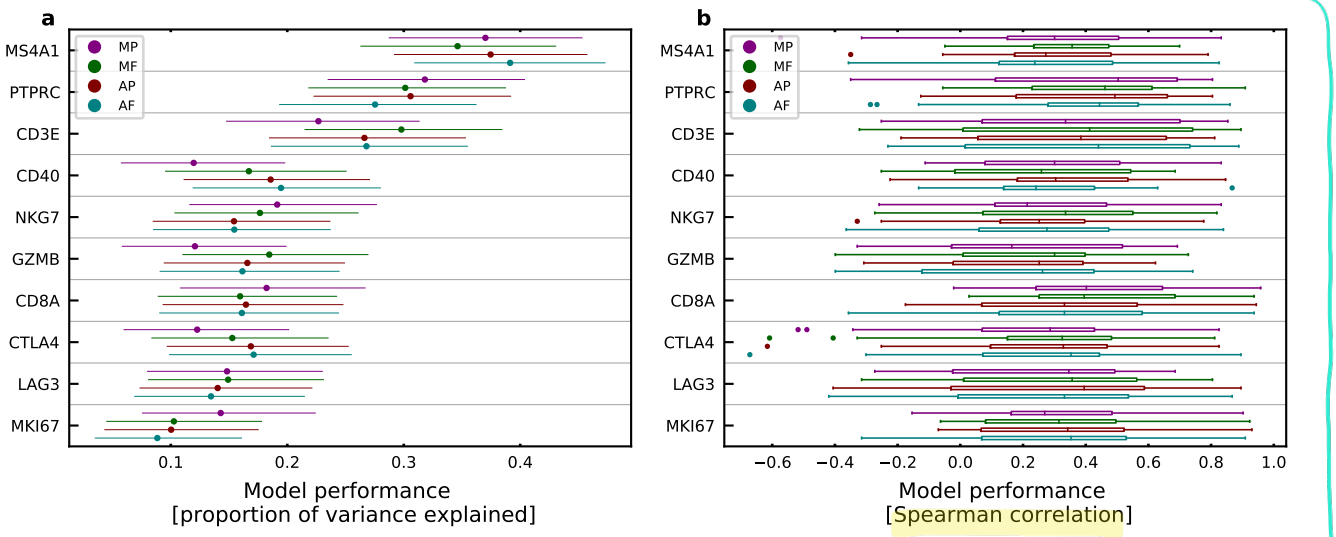

Figure 3. Spatial transcriptomics results for the ten transcripts with the highest proportion of variance explained. a) proportion of variance predicted (as modelled by the fixed effects in the LME model with transcripts as dots and bars that indicate the 95% confidence intervals). b) distribution of Spearman correlations between model predictions and spatially estimated gene expression values for each model and transcript. Boxplots were generated analogously to Figure 1.

of variance predicted between the prediction models is generally small compared to the difference between transcripts.

4. Discussion

In this study, we compared four models that directly predict set-level regression labels for a set of images, both in an MNIST simulation as well as with the prediction of gene expression from WSIs of H&E stained breast tumor sections. All of these models differ in their assumptions regarding the relationship between the set-level label and the individual instances. Two of the models that we investigated use attention mechanisms to generate slide-level predictions based on a set of tile representations. To the best of our knowledge, this is the first application of these attention mechanisms to regression objectives in the histopathology domain, where they can be used to allow for varying contributions of individual WSI tiles to the WSI-level label.

The MNIST simulation experiment indicates that in some scenarios, where relatively few images contribute to the set-level label, the flexibility of allowing for varying contributions of instances to the set-level label allows models with attention mechanisms to better capture the relationship between set and instance labels. For a high proportion of noise labels, the weak label model MP is inferior to the other three models. This is plausible because this model weights all predictions equally, including predictions of noise images which dominate the predictions of the MP model for high proportions of noise.

However, when predicting gene expression from WSIs, our results indicate that using an attention mechanisms leads to a similar model performance as the common approach of using slide-level labels as a weak label for individual tiles. This could be the case for several reasons. If the MNIST simulation is a somewhat reasonable model of predicting gene expression from WSIs with regards to the contributions of image tiles to the WSI-level label, the percentage of image tiles that does not contribute to the WSI label may be in the range between 25% and 50% for our histopathology application. Another possibility is that the MNIST simulation is not a sufficiently realistic model for the histopathology application and that the observations from this experiment are therefore not or only poorly transferable. It is indeed unlikely that some tiles do not contribute to the WSI-level gene expression label at all, it appears more plausible that the difference of contributions of different WSI regions is more gradual. While the MF model performance is comparable to the MP model in the MNIST simulation, we find it to be inferior to the other models in the histopathology setting. This may be due to the large difference in set sizes between the MNIST experiment and WSIs. While the mean of features on a relatively small MNIST set of 32 images may not result in a loss of information, this may not translate well to averaging the features of thousands of tiles of a WSI. Nevertheless, the difference in prediction performance (median Spearman correlations) is quite small, albeit statistically significant with $p < 0.05$. The prediction performance (median Spearman correlation) of the 174 transcripts investigated is lower in

the external test set as compared to the internal set for all models. This may be due to differences in sample preparation between the tissue material in the data sets that leads to diminished generalization. The decrease in prediction performance (median Spearman correlation) is statistically significant and larger for models with an attention mechanism. This may be because slightly worse generalisation to external data could potentiate itself through the attention mechanism. Nevertheless, we conclude that all models generalize relatively well to the external test set, as the value of the differences in correlation is rather small.

We applied spatial transcriptomics profiling to evaluate the performance in spatial expression predictions by the four different models, which did not reveal apparent differences in the ability to predict local expression. While the LME model analysis accounts for slide level differences (modelled as a random effect) there are other sources of variability that are not accounted for here, including noise due to image registration. There is also an expected upper bound on how well expression can be predicted since there has to be a morphological phenotype associated with each transcript. While the ST analysis does unfortunately not allow to draw any new conclusions, the findings also do not contradict the findings from the MNIST simulation or bulk RNA-seq analysis. While the prediction performance (Spearman correlation) for some slides are relatively high across transcripts and models, the proportions of variance predicted (as modelled by the LME model) is only relatively high for a few transcripts. Both a larger sample size than 22 WSIs, as well as potentially improved registration may be necessary to observe differences between the investigated models, if they exist. This could be due to relatively poor local prediction performance or limitations of the analysis that we conducted. Spatial transcriptomics unfortunately remains costly and challenging to implement at scale. However, it is to be expected that the availability will increase widely, which may provide opportunity to revisit the research questions that we investigated in this study.

This study has several limitations. The MNIST example may oversimplify the histopathology setting, and it is therefore difficult to ultimately assess its usefulness beyond a proof of concept. Furthermore, predicting gene expression from WSIs is a challenging task. The upper limits to the Spearman correlations and numbers of genes that are statistically significantly predictable are similar in in several studies that all use different modelling approaches [6, 18, 21, 22]. This upper limit may be due to differences in tissue used for RNA sequencing and scanning, as well as limits of the association between morphologies in WSIs of H&E stained tissue sections and gene expression. This limit in the currently reported prediction performances may obfuscate differences between the modelling approaches that we investigated. This applies especially to the spatial transcriptomics data, where the correlation is further limited by the accuracy of the registration between the sequenced and scanned tissue sections.

5. Conclusion

This study indicates that while there is some evidence that the MF model is inferior compared to the other three models investigated, the performance differences are relatively minor. This may also be seen as encouraging, as it could indicate that researchers are not at a large risk of poor model performance due to a specific model choice. This may also be supported by the similarity of results between prior studies that predicted gene expression from WSIs despite their differences in modelling choices. As opposed to recent studies focusing on MIL classification objectives in computational pathology, the attention mechanism that we investigated does not seem to provide a strong benefit in model performance as compared to using slide-level labels as weak labels for all image tiles in the regression setting. Furthermore, the generalizability appeared to be poorer for models with an attention mechanism. Using weak labels, as in the MP model, appears robust, has fewer model parameters and adds less complexity to the already complex domain of computational pathology. Worth noting is that the MP model appears to perform poorly if only a small fraction of images contributes to the set-level label, as demonstrated in the MNIST simulation. We conclude that while the flexibility of attention mechanisms warrants further investigation, based on current results for the regression problem of gene expression prediction, the commonly used weak label approach might be preferable.

6. Acknowledgements

This project was supported by funding from the Swedish Research Council under the frame of ERA PerMed(ERAPERMED2019-224 - ABCAP), Swedish Research Council, Swedish Cancer Society, Karolinska Institutet (Cancer Research KI, StratCan), MedTechLabs, Swedish e-science Research Centre (SeRC), Stockholm Region, Stockholm Cancer Society, Swedish Breast Cancer Association. The authors would like to acknowledge patients, clinicians, and hospital staff participating in the SCAN-B study, the staff at the central SCAN-B laboratory at Division of Oncology, Lund University, the Swedish national breast cancer quality registry (NKBC), Regional Cancer Center South, and the South Swedish Breast Cancer Group (SSBCG). We thank the TCGA Research Network.

References

[1] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. Fitting linear Mixed-Effects models using lme4. *Journal of Statistical Software, Articles*, 67(1):1–48, 2015. 5

[2] Y Benjamini and Y Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc.*, 1995. 5

[3] Christian Brueffer, Johan Vallon-Christersson, Dorthe Grabau, Anna Ehinger, Jari Häkkinen, Cecilia Hegardt, Janne Malina, Yilun Chen, Pär-Ola Bendahl, Jonas Manjer, Martin Malmberg, Christer Larsson, Niklas Loman, Lisa Rydén, Åke Borg, and Lao H Saal. Clinical value of RNA Sequencing–Based classifiers for prediction of the five conventional breast cancer biomarkers: A report from the Population-Based multicenter sweden cancerome analysis Network—Breast initiative. *JCO Precision Oncology*, 2(2):1–18, Nov. 2018. 4

[4] Gabriele Campanella, Matthew G Hanna, Luke Geneslaw, Allen Miraflor, Vitor Werneck Krauss Silva, Klaus J Busam, Edi Brogi, Victor E Reuter, David S Klimstra, and Thomas J Fuchs. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nat. Med.*, 25(8):1301–1309, Aug. 2019. 1

[5] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, Oct. 2012. 4

[6] Yu Fu, Alexander W Jung, Ramon Viñas Torne, Santiago Gonzalez, Harald Vöhringer, Artem Shmatko, Lucy R Yates, Mercedes Jimenez-Linan, Luiza Moore, and Moritz Gerstung. Pan-cancer computational histopathology reveals mutations, tumor composition and prognosis. *Nature Cancer*, 1(8):800–810, Aug. 2020. 2, 8

[7] K He, X Zhang, S Ren, and J Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, June 2016. 3

[8] Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2127–2136. PMLR, 10–15 Jul 2018. 2, 3

[9] Paul Cd Johnson. Extension of nakagawa & schielzeth's R2GLMM to random slopes models. *Methods Ecol. Evol.*, 5(9):944–946, Sept. 2014. 5

[10] Y Lecun, L Bottou, Y Bengio, and P Haffner. Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324, Nov. 1998. 3

[11] Ming Y Lu, Tiffany Y Chen, Drew F K Williamson, Melissa Zhao, Maha Shady, Jana Lipkova, and Faisal Mahmood. AI-based pathology predicts origins for cancers of unknown primary. *Nature*, May 2021. 2

[12] Ming Y Lu, Drew F K Williamson, Tiffany Y Chen, Richard J Chen, Matteo Barbieri, and Faisal Mahmood. Data-efficient and weakly supervised computational pathology on whole-slide images. *Nat Biomed Eng*, Mar. 2021. 2, 3

[13] M Macenko, M Niethammer, J S Marron, D Borland, J T Woosley, Xiaojun Guan, C Schmitt, and N E Thomas. A method for normalizing histology slides for quantitative analysis. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1107–1110, June 2009. 4

[14] N Otsu. A threshold selection method from Gray-Level histograms. *IEEE Trans. Syst. Man Cybern.*, 9(1):62–66, Jan. 1979. 4

[15] Hans Pinckaers, Bram van Ginneken, and Geert Litjens. Streaming convolutional neural networks for end-to-end learning with multi-megapixel images. *IEEE Trans. Pattern Anal. Mach. Intell.*, PP, Aug. 2020. 1

[16] PyTorch MNIST Tutorial. https://pytorch.org/tutorials/beginner/former_torchies/nnft_tutorial.html. Accessed: 2021-07-19. 3

[17] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, Dec. 2015. 3

[18] 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, Aug. 2020. 2, 8

[19] David Tellez, Geert Litjens, Jeroen van der Laak, and Francesco Ciompi. Neural image compression for gigapixel histopathology image analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2):567–578, Feb. 2021. 1

[20] Mei Wang, Daniel Klevebring, Johan Lindberg, Kamila Czene, Henrik Grönberg, and Mattias Rantalainen. Determining breast cancer histological grade from RNA-sequencing data. *Breast Cancer Res.*, 18(1):48, May 2016. 4

[21] Yinxi Wang, Kimmo Kartasalo, Philippe Weitz, Balazs Acs, Masi Valkonen, Christer Larsson, Pekka Ruusuvuori, Johan Hartman, and Mattias Rantalainen. Predicting molecular phenotypes from histopathology images: a transcriptome-wide expression-morphology analysis in breast cancer. *Cancer Res.*, Aug. 2021. 2, 4, 8

[22] Philippe Weitz, Yinxi Wang, Kimmo Kartasalo, Lars Egevad, Johan Lindberg, Henrik Grönberg, Martin Eklund, and Mattias Rantalainen. Transcriptome-wide prediction of prostate cancer gene expression from histopathology images using co-expression based convolutional neural networks. *CoRR*, abs/2104.09310, 2021. 2, 8