# AI-Driven Biomarker localisation Prediction from H&E Stained Tissue.

Idris Iritas (Student 2743506) Hugo Horlings (Supervisor)
Rolf Harkes (Daily Supervisor) Marten Postma (Examiner)

June 24, 2024

Major research project Bioinformatics & Systems Biology (XM_0103, 60EC)

# Abstract

In order to provide a prognosis for breast cancer patients, pathologists often spend significant time annotating H&E-stained whole-slide images, a time-consuming process constrained by the limited number of expert pathologists. Tumour-infiltrating lymphocytes (TILs) are crucial indicators in cancer prognosis and therapy response. Automating TIL identification could speed up biopsy prognoses. To facilitate this, many efforts focus on automating slide annotation through deep-learning methods, which require vast amounts of annotated data that are not readily available. A novel method called co-detection by indexing (CODEX) enables imaging of up to 60 bio-markers in a single image and allows for creating a subsequent H&E scan of the same tissue slide. In this study we develop a method to create binary segmentation labels from CODEX data through image registration, H&E patch extraction, and binary mask generation. We aim to answer the research question: Can a U-Net model accurately predict bio-marker presence in H&E-stained histopathological breast cancer images using CODEX-derived training data, and can this model accurately identify lymphocytes in direct contact with tumour cells? Our results show promising performance for two bio-markers, accurately predicting DAPI signal presence and epithelial tissue with F1 scores of 0.718 and 0.769, respectively, after training on two whole-slide images. However, the model performs poorly on several under-represented bio-markers, indicating key areas for future optimization. We observed the model's ability to identify TIL regions, although output probabilities are not high enough to consistently distinguish positive from negative signals. By enhancing our training data quality and quantity through improved patch extraction, mask generation, and class-imbalance handling methods, we aim to boost model performance, robustness, and output confidence. These developments could enable the deployment of the model in a medical context, automating classification tasks for expert pathologists and allowing them to focus more on quality control.

# Contents

# 1 Introduction

## 1.1 Motivation & Research question

Breast cancer has been determined to be the most frequently diagnosed cancer in women (Łukasiewicz et al. 2021), with an estimated 2.3 million new cases and over 660,000 deaths in 2022 (Ferlay et al. 2018). Prior studies have shown early-stage screening and diagnosis of breast cancer plays a key role in lowering the mortality rate (Broeders et al. 2012). When a positive diagnosis is concluded, clinical pathologist experts are tasked with assessing tumour biopsies for subsequent treatment plans. This assessment is a very time-consuming task which may delay patient treatment prescription. The field of pathology has seen a rise in automation and computational applications in recent years, in part due to the digitization of tumour microscopy slides amidst the COVID-19 pandemic in 2020 (Schwen et al. 2023; Cui and D. Y. Zhang 2021). This digitization of pathology data allows for automation of high-throughput cancer screening, increasing speed and potentially efficiency of slide annotation. Training these AI models requires massive amounts of annotated data, though the aforementioned shortage of pathologists and the time needed to properly annotate these cells limits the amount of data available. For this reason, novel methods are necessary for labeled data generation to facilitate the training of better classification models. Successful implementation of deep-learning based AI models may help facilitate faster treatment amidst a global shortage of specialized pathologists (A. S. Leong, Joel, and W.-M. Leong 2005), as well as aid in predicting survival and treatment outcomes.

Pathologists commonly assess biopsies on Haematoxylin & Eosin (H&E)-stained whole-tissue slides. These slides contain cross-sections of cancerous tissue biopsies stained with haematoxylin to stain the cell nuclei and other basophilic organelles, and eosin to stain most of the extracellular matrix as well as other eosinophilic cell components and organelles (Chan 2014). This staining is the golden standard for pathology data, and is used in pathology laboratories worldwide. For this reason, specialized AI models are often trained on H&E-stained slide images to carry out diverse ranges of classification tasks. H&E slides are commonly processed into two distinct formats: Whole slide images (WSIs) and Tissue micro-arrays (TMAs). They both have advantages and drawbacks when it comes to analysis of these biopsies. WSIs contain a large number of cells, and thus a large amount of data. The drawback of these slides is that there is no heterogeneity, as the entire tissue slide comes from the same biopsy. TMAs are a collection of many small biopsy samples from different patients. These arrays contribute to tumor heterogeneity by providing data of many different tumour varieties and tumour environments, facilitating improved model generalization in the field of computational pathology.

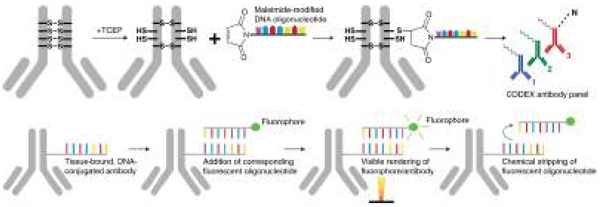
+TCEP
HS
SH
S
N
O
Maleimide-modified DNA oligonucleotide
1
2
3
N
CODEX antibody panel
Fluorophore
Fluorophore
Tissue-bound, DNA-conjugated antibody
Addition of corresponding fluorescent oligonucleotide
Visible rendering of fluorophore/antibody
Chemical stripping of fluorescent oligonucleotide

**Figure 1: Visual representation of CODEX workflow.** Maleimide-DNA oligonucleotides are conjugated to IgG antibody, consequently fluorophore-tagged oligonucleotides bind to antibody complexes and emit signals. Finally, oligonucleotides are chemically stripped. (Image from Black et al. 2021)

Our proposed method of obtaining accurate ground-truth training labels without the need for annotations from expert pathologists involves a method called Co-detection by indexing (CODEX). CODEX is a high-dimensional multiplex bio-imaging method which involves conjugating specific maleimide-modified DNA sequences to biomarker-specific antibodies and subsequently hybridizing DNA-conjugated fluorophores to these biomarker-antibody complexes (Figure 1) (Black et al. 2021).

This method allows for multiplexed imaging of a wide range of bio-markers, while leaving the tissue undamaged which allows for subsequent H&E staining and bright-field imaging. In summary, a whole tissue slide is incubated with a panel of primary IgG antibodies that bind to specific biomarkers to be analysed. These primary antibodies are conjugated to DNA-tags specific for one of several fluorophores. In several cycles, DNA-tagged antibodies are applied, hybridized to complementary DNA-tagged fluorophores and fluorescence imaging is performed. After every cycle, the conjugated fluorophores are chemically stripped and a new set of fluorophores is applied and analysed. These steps repeat until all biomarkers are imaged. After analysis, the output is an overlapping multi-channel image where each channel displays the captured signal of one biomarker. This methodology allows for analysis of a wide range of biomarkers within a single sample. Multiplexed image analysis on tumour biopsies provides ground truth labels on a single-cell level in WSIs and TMAs, whereas pathologists aim to predict cell types solely from H&E-stained images. Prior research has focused on training AI models on ground-truth cell labels provided by pathologist consensus, spatial arrangement of cells and immuno-staining (Couture 2022). However, these labels are often susceptible to human bias in the case of pathologist consensus, may lack sufficient information individually, or do not provide labels of adequate quality and quantity for robust model training.

Experiments carried out by Y. van Cleef, a prior student working in this research group, demonstrated that a model trained on CODEX ground-truth labels outperforms a model trained on labels obtained through triple-staining experiments, in part due to better alignment of data and label images [1]. The aim of this study is to train a deep-learning based segmentation model using CODEX-derived binary masks in order to predict the presence and absence of several bio-markers from H&E image data.

To achieve this, CODEX experimental data utilizing a diverse panel of antibodies is processed to generate binary masks per individual marker. These binary masks are registered onto the subsequently imaged H&E scan and serve as the gold standard labels for training a segmentation model. This study aims to answer the following 2 questions:

1. *Can a U-Net model accurately predict biomarker presence in H&E-stained histopathological breast cancer images using CODEX-derived training data?*
2. *Can our trained U-Net model accurately identify lymphocytes in direct contact with tumor cells in histopathological breast cancer images?*

The ability of our model to predict bio-marker image masks will be assessed through analysis of model performance metrics, whereas the number of lymphocytes neighbouring tumour cells can be compared between ground-truth data and model output.

[1] Training an AI-based single cell classifier using different sources of ground truth data - Ynze van Cleef (Student ID: 2650806), 2023

## 1.2 State of the art

The field of pathology has seen many advancements in recent years, making way for the field of "computational pathology". Many years of research into molecular prognostic bio-markers in cancer have lead to an increasing complexity in cancer prognosis (Echle et al. 2021). Echle *et al.* states that deep-learning systems can approach human performance in tumour detection, grading and subtyping, and that AI-based methods are able to extract high-level features from H&E data that are unfeasible for humans to predict.

### 1.2.1 Recent advancements in AI

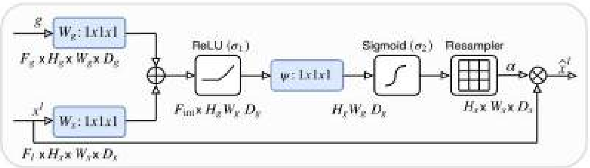
$g$
$W_g$: 1x1x1
$F_g \times H_g \times W_g \times D_g$
$x^l$
$W_x$: 1x1x1
$F_l \times H_x \times W_x \times D_x$
ReLU ($\sigma_1$)
$F_{int} \times H_g W_g D_g$
$\psi$: 1x1x1
Sigmoid ($\sigma_2$)
$H_g W_g D_g$
Resampler
$H_x \times W_x \times D_x$
$\alpha$
$\hat{x}^l$

**Figure 2: Schematic overview of information flow in the attention gate architecture.** During the decoding process, the attention gate receives input from the up-scaling ($x^l$) and skip connection ($g$) layer. Through a range of learnable weights and activation functions, the attention gate emphasizes or suppresses features from both input sources enhancing the model's ability to accurately segment target objects in the image. Finally, the re-sampler increases the dimensions through a resizing algorithm. (Image from Oktay et al. 2018)

The ability for deep-learning systems to extract more information from data reliably is facilitated by recent advancements in the field of deep-learning. One major recent advancement in the field of AI is the transformer architecture, introducing the self-attention mechanism (Vaswani et al. 2017) which has been implemented into many models including U-NETs for applications in the field of computer vision and computational biology. This advancement has led to the development of the attention gate architecture (Figure 2) which allowed for self-attention in computer vision tasks (Oktay et al. 2018).

### 1.2.2 U-Net architecture

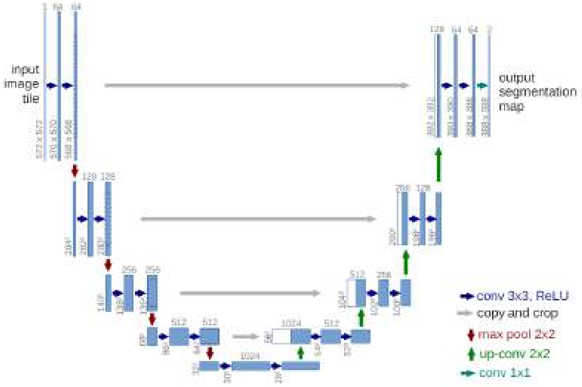

**Figure 3: Schematic representation of the original U-NET architecture.** Input layer undergoes convolution and pooling through the encoder. Decoder up-scales the image alongside skip connections, subsequent convolutions processes signal into binary segmentation map. (Image from Ronneberger et al. 2015)

The U-Net architecture, first proposed in 2015 (Ronneberger, Fischer, and Brox 2015) is a model architecture developed for biomedical image segmentation. This model architecture consists of an input layer followed by multiple convolution and pooling layers like a traditional convolution neural network. After these encoding layers, the network upscales the data again, while resized outputs of the prior encoder layers serve as spatial information for the up-scaling process. These resized outputs are named "skip connections". After the decoder up-scales the data again, several convolutions are applied again to transform the data into a binary classification image (Figure 3).

After the development of attention gates, Oktay *et al.* (Oktay et al. 2018) presented the first implementation of attention mechanisms into the U-Net architecture, displaying a general improvement in model accuracy. Since then, many U-NET architectures have been developed for specific tasks and many benchmarks have been performed to compare the performance of a range of architectures on identical datasets (Bhandary et al. 2023). Training U-NETs on histological data for segmentation tasks has been documented before, for example by Bulten *et al.* (Bulten et al. 2019), but to our knowledge prediction of multiplexed masks through U-NETs has not been described prior to this research.

### 1.2.3 Advancements in computational pathology

Song *et al.* (Song et al. 2023) presents an extensive overview into recent and prospective developments in the field of computational pathology. One issue described in this overview is the limitation of tiling WSIs. When a tissue slide is divided into smaller tiles, a model trained on these tiles may miss out on spatial information surrounding the tile. Implementation of spatial information through graph-neural-networks is a very recent development, and shows promising results in improving model accuracy (Ahmedt-Aristizabal et al. 2022). Other types of models often found in recent literature are foundation models, which often combine Natural-Language-Processing, large model architectures and massive amounts of data in order to generalize well on a large variety of tasks within a specific field (Lu et al. 2023). Not all of these implementations are feasible within this research project, but they serve as possible directions for subsequent research.

### 1.2.4 Tumour-infiltrating lymphocytes

Tumour-infiltrating lymphocytes (TILs) are prominent markers for immune response in tissue samples and have been shown to be a moderate prognostic marker for survival outcomes (Gooden et al. 2011). TILs consist of various immune cells, including T cells, B cells, and natural killer cells, which migrate into the tumour micro-environment in response to tumour antigens. The presence and density of TILs within tumours have been associated with improved prognosis and better clinical outcomes in several types of cancer, including breast cancer, testicular cancer, colorectal cancer and melanoma (Pagès et al. 2009; Underwood 1974). Additionally, through a meta-analysis study, Hu *et al.* shows that tumour infiltration of CD45RO-positive T-lymphocytes predicts favourable prognosis in solid tumours (Hu and Wang 2017).

The evaluation of TILs in histopathological samples has become increasingly important in the context of immuno-oncology. Immunotherapy, which leverages the body's immune system to fight cancer, has revolutionized cancer treatment, and the assessment of TILs can provide critical insights into the immune landscape of tumours (Clemente et al. 1996). High levels of TILs have been associated with increased response rates to checkpoint inhibitors, such as anti-PD-1 and anti-CTLA-4 therapies, highlighting their potential role as predictive biomarkers for immunotherapy response (Hendry et al. 2017).

For this reason, the ability to measure the number and distribution of TILs within a tissue slide is valuable for prognosis and provides important information when prescribing treatment to patients. Recent advancements in digital pathology and image analysis have facilitated automatic quantification and spatial analysis of TILs, enabling more precise and reproducible assessments (Saltz et al. 2018). High-level features such as the spatial distribution of lymphocytes and tumour cells could be indicative factors for lower-level features, such as immunity-specific pathways for which treatments are available.

Sobral *et al.* describes a correlation

between tumour infiltration of CD8+ T-lymphocytes and mutations in PIK3CA, a protein involved in the PI3K pathway, which is commonly altered in ER+ breast cancer (Sobral-Leite et al. 2019). The PI3K pathway is critical for cell growth, survival, and metabolism, and its dysregulation is frequently observed in various cancers. Mutations in PIK3CA can lead to the activation of downstream signaling pathways that promote tumour growth and survival. The presence of CD8+ T-lymphocytes, which are cytotoxic T cells, in PIK3CA-mutant tumours suggests a potential interplay between oncogenic signaling pathways and the immune microenvironment.

Understanding these correlations may provide valuable insights for developing targeted therapies that not only inhibit tumour growth but also enhance anti-tumour immune responses. Robust methods to identify and potentially classify TILs may lead to more effective and personalized treatment strategies without requiring expensive and time-consuming data acquisition from tumour biopsies.

# 2 Methodology

This section outlines the image-processing workflow, as well as the data preparation for model-training. The general workflow is outlined in Figure 4. Tumour biopsies are processed into tissue and tumour micro-array microscopy slides, which then undergo CODEX multiplexed image analysis followed by a subsequent H&E staining and scan. Raw scan data is processed into background-normalized readable image formats. Multiplexed and H&E images are overlaid through image registration protocols and image masks are generated from multiplexed data through thresholding and filtering. Finally, images are separated into tiles which are fed to a deep-learning U-NET model. Model inference output is utilized for classification tasks on a cellular level.

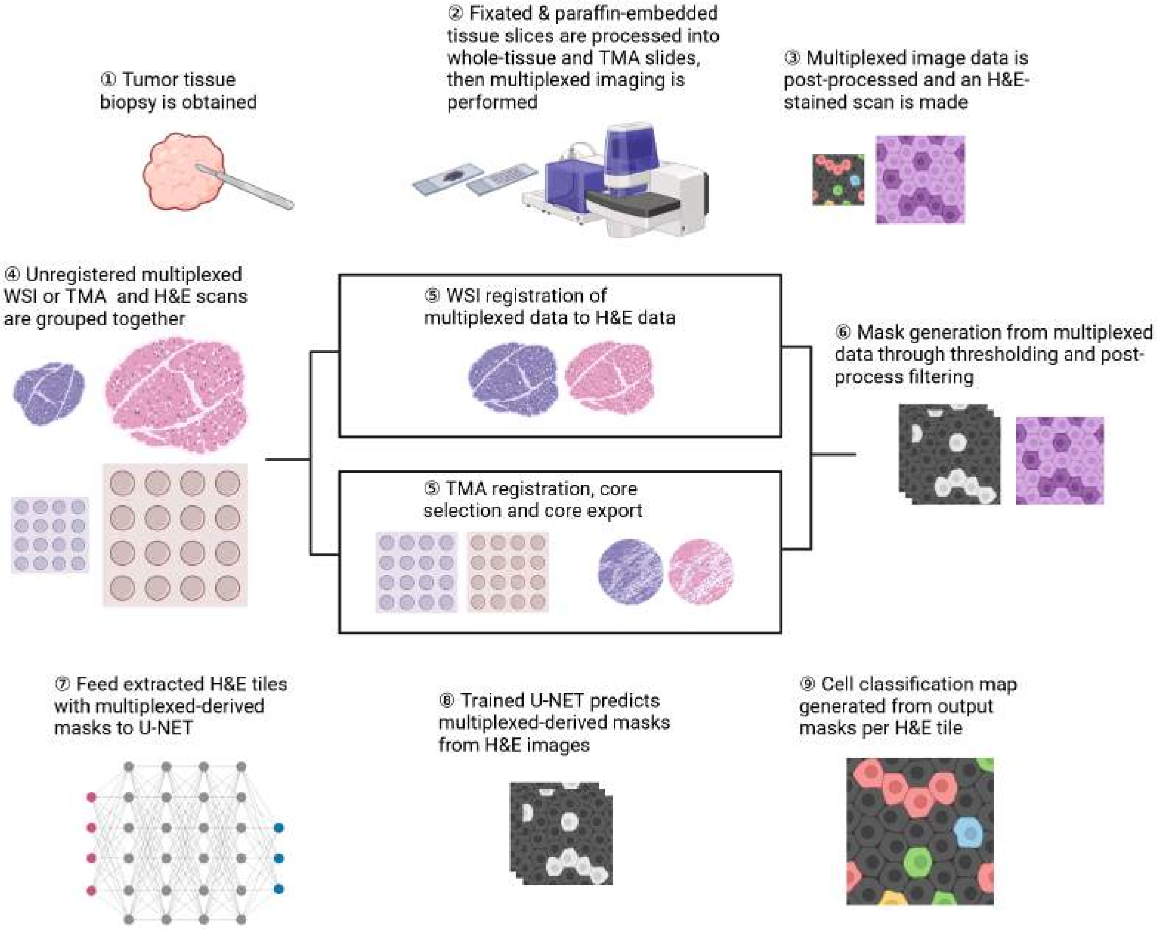
① Tumor tissue biopsy is obtained
② Fixated & paraffin-embedded tissue slices are processed into whole-tissue and TMA slides, then multiplexed imaging is performed
③ Multiplexed image data is post-processed and an H&E-stained scan is made
④ Unregistered multiplexed WSI or TMA and H&E scans are grouped together
⑤ WSI registration of multiplexed data to H&E data
⑤ TMA registration, core selection and core export
⑥ Mask generation from multiplexed data through thresholding and post-process filtering
⑦ Feed extracted H&E tiles with multiplexed-derived masks to U-NET
⑧ Trained U-NET predicts multiplexed-derived masks from H&E images
⑨ Cell classification map generated from output masks per H&E tile

**Figure 4: Schematic overview of research workflow.** Tissue samples are obtained and processed into microscopy slides as whole-slide images or tumour micro-arrays. CODEX multiplexed imaging analysis is performed and captured images are post-processed into a usable format with subtracted background intensity. Multiplexed images are registered to H&E scans and further cropped in the case of micro-arrays. Masks are generated through intensity thresholding and post-processed for noise removal when applicable. Finally, H&E and corresponding mask tiles are fed to a U-NET model for training, and inference on H&E images will produce marker-specific masks used for downstream cell classification tasks. Figure created with BioRender.com.

## 2.1 H&E and CODEX data

Original CODEX-data was composed of 27 channels consisting of various bio-marker signals (figure s1). Following a thorough evaluation of the measured signals across multiple channels by a trained pathologist, all but 7 channels were excluded from the data due to noisy signals, signal bleed-through from previous measurements and potential nonspecific signals on certain biomarkers. Table 1 displays the biomarkers that were retained following this evaluation.

**Table 1: CODEX-panel markers, their functions and cellular localisation.** Markers are limited to those included after pathologist examination.

| Marker | Displays | Localisation |
|---|---|---|
| DAPI | General nuclei | Nucleus |
| CD163 | Monocytes & macrophages | Membrane/Cytoplasm |
| CD45RO | Activated lymphocytes | Membrane/Cytoplasm |
| EpCAM | Epithelial adhesion glycoprotein | Membrane |
| PanCK | General epithelial cells | Membrane |
| E-CAD | Epithelial adhesion & signal transduction | Membrane |
| SMA | Smooth muscle tissue | Cytoplasm |

In the context of this research, multiplexed analysis and H&E staining and scanning are performed in advance and provided for subsequent image processing. In short, multiplexed images are captured by the Phenocycler and Phenoimager fusion (Akoya) with a 40x objective lens using an excitation and exposure time of 150 ms, with the exception of PD-L1, which was excited and captured for 200 ms, and CD163, CD66b and CD138 which were excited and captured for 300 ms. Captured images were stored at a bit-depth of 14 and subsequently saved as 16-bit multi-channel images. After fluorescence imaging, H&E scans were made of the same tissue slide by automatic staining using the Tissue-Tek Prisma Plus (Sakura) and subsequent brightfield imaging, using the PANNORAMIC 1000 (3DHistech).

## 2.2 Image registration

In order to perform image registration of codex data onto H&E WSI-imaging data, raw H&E image data was initially acquired in .mrxs format, while raw codex data was derived from intermediate raw data$^{B.2.1}$. To prepare the data for registration, both images are loaded into QuPath (v0.5.1) (Bankhead et al. 2017) using their respective compatible image servers (OpenSlide for H&E and Bio-Formats for CODEX). H&E images were cropped and exported as 8-bit pyramidal .ome.tif$^{B.2.2}$. For subsequent registration, the DAPI channel was extracted from the CODEX data and exported as 8-bit pyramidal .ome.tif$^{B.2.3}$.

Image registration of whole-slide images was performed using the Fiji distribution of ImageJ (v1.54f) (Schindelin et al. 2012) in conjunction with the Warpy extension (Chiaruttini et al. 2022) powered by ElastiX (v5.0.1) (Klein et al. 2009). In order to use Warpy, a QuPath dataset was created in ImageJ using the BigDataViewer plugin (v10.4.14) (Pietzsch et al. 2015). The 8-bit H&E scan served as the fixed source, while the extracted 8-bit DAPI channel was utilized as the moving source. This process involved creating thin-plate spline deformations in a pre-assigned grid formation over the entire image, where a patch size parameter of 1000 microns with no spacing was chosen. Warpy and Elastix then optimize the dice coefficient between the 2 images for each patch through stochastic gradient descent. After registration, a transform_i_j.json file was output, where i is the moving- and j is the fixed-source image ID assigned by QuPath.

In order to apply this transformation to the entire 16-bit CODEX dataset, the QuPath IDs of the 8-bit H&E image and 16-bit CODEX image were identified and the transform.json was renamed to include appropriate moving- and fixed-source IDs. In order for QuPath to recognize the registration, the transform .json file was copied to the directory associated with the identifier of the moving CODEX source within the QuPath project's data directory.

In QuPath, the 8-bit H&E image was accessed and utilizing the Warpy Image Combiner plugin for QuPath (Chiaruttini et al. 2022), the 16-bit CODEX data was selected for registration. The "Warpy option was selected resulting in a registered concatenated image server. Finally, a range of registered channels corresponding to the markers shown in table 1 was exported in 16-bit pyramidal .ome.tif format$^{B.2.4}$.

## 2.3 Data & label preparation

In order to obtain ground-truth labels for model training, image masks were generated from the registered multiplexed image data through iterative manual threshold selection guided visually by a trained pathologist. Manual threshold selection was necessary due to errors in automated thresholding metrics caused by artefacts present in the CODEX data. Thresholded images were de-noised through a 4x4 median filter and exported as whole-slide image masks through OpenCV-integrated QuPath scripting$^{B.2.5}$ (Bradski 2000).

Given the limited quantity of data, identical images and masks were pre-tiled at a native resolution of 0.25 µm/pixel into 512 by 512 pixel tiles. These tiles were then divided into training, testing, and validation sets with a distribution of 80%, 10%, and 10% respectively. This method contrasts with the traditional approach of dividing whole-slide images into different sets by assigning entire slides to each set. Tiling was performed through scripting$^{B.2.6}$ in conjunction with DLUP (v0.3.27), OpenSlide (v4.0.0.2) and Libvips (v8.15.2) (Teuwen et al. 2024; Goode et al. 2013; Martinez and Cupitt 2005).

In order to address class imbalance during model training, a subset of the dataset was created for each model except for the positive control (DAPI). Specifically, 90% of the images in this subset contained at least 0.5% positive pixels in the image mask. This approach was implemented to prevent the model from becoming biased towards the majority negative class, which can lead to the model getting stuck in a local minimum. By ensuring a more balanced representation of classes, the training process was optimized for better performance and generalization.

## 2.4 Model training

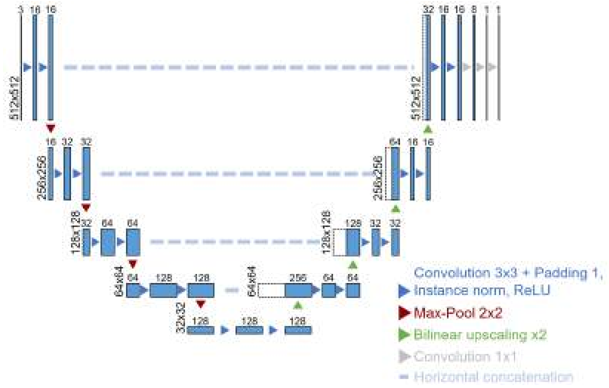

**Figure 5: Schematic representation of the U-Net implementation used in this research.** Each tensor is labeled with image dimensions on the left and the number of features on top. Operations between tensors are depicted by colored arrows, with each color representing a specific operation as detailed in the legend. Skip connections are shown as dashed lines, and concatenated tensors are indicated with a dashed border.

A variety of U-Net implementations were explored to determine the most effective architecture for this segmentation tasks. After evaluating several options, a U-Net model inspired by the one used by Facebook for the fastMRI challenge was selected[2,3]. This specific implementation was chosen due to its demonstrated performance and efficiency on our segmentation dataset. A schematic overview of the specific model architecture used can be found in figure 5

In order to train the model, the PyTorch (v2.3.0)(Paszke et al. 2019) framework was used in conjunction with the CUDA 12.1(Nickolls et al. 2008) compute platform for GPU-accelerated training. All models were trained on 3x512x512 RGB images and 1x512x512 binary masks using a batch size of 8. To enhance model generalization, image augmentations were applied and categorized into two groups: geometric transformations and color transformations. Geometric transformations, consisting of vertical flips, 90-degree factor rotations, and random -45 to 45 degree rotations were applied to both input images and labels. On the other hand, color transformations, consisting of Gaussian blur and HED jitter(Tellez et al. 2018)[4], were exclusively applied to the input images.

In order to compute the loss between output and label tensors, the PyTorch implementation of Binary Cross-entropy with logits loss was utilized. This loss function is defined as shown in equation 1:

$$\ell(\hat{\mathbf{Y}}, \mathbf{Y}) = -(\mathbf{Y} \odot \ln(\sigma(\hat{\mathbf{Y}})) + (1 - \mathbf{Y}) \odot \ln\left(1 - \sigma\ (\hat{\mathbf{Y}})\right)) \quad (1)$$

Where $\sigma(x)$ is the sigmoid activation function. Here, $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ represent the model's predicted outputs and true labels respectively. When calculating the loss over a batch of samples, the loss values are averaged over all samples in the batch. The sigmoid activation function, defined as $\sigma\left(x\right) = \frac{1}{1+e^{-x}}$, is applied to the model's output to scale the raw logits into a probability range of [0, 1]. This allows the model to output probabilities instead of raw logits, facilitating convenient inference processing. By using sigmoid activation in conjunction with binary cross-

[2] github.com/NKI-AI/kandinsky-calibration/blob/main/src/models/components/unet.py (51deb1b)

[3] github.com/facebookresearch/fastMRI/blob/main/fastmri/models/unet.py (cc42c7a)

[4] github.com/gatsby2016/Augmentation-PyTorch-Transforms (1fbfcf9)

entropy loss, we mitigate potential vanishing gradients during training.

As a parameter optimizer, the Adam algorithm(Kingma and Ba 2014) was chosen for its established efficiency in segmentation tasks(Mortazi et al. 2023), documented implementation in the PyTorch framework and demonstrated performance on the dataset used for this study. For parameter optimization, the Adam algorithm was used with default parameters ($\gamma = 0.001$, $\beta_1 = 0.9$, $\beta_1 = 0.999$, $\epsilon = 10^{-8}$). A validation set was utilized to monitor validation-loss during training in order to utilize early-stopping, aiding in preventing the model from over-fitting to the training set. The validation set was further utilized for optimizing hyper-parameter tuning, unbiased inference-processing and regularization techniques such as image-augmentation during model development.

## 2.5 Model evaluation

In order to assess performance of the trained segmentation models, a proper threshold for the inference output had to be determined beforehand. In order to compute these optimal thresholds, $F_1$ scores were maximized on the validation set per model. To subsequently assess the performance of each model, the following metrics were computed at respective optimal validation thresholds applied to the test set predictions to evaluate the models' effectiveness across different aspects of image segmentation: Dice similarity coefficient ($\frac{2|\hat{\mathbf{Y}} \cap \mathbf{Y}|}{|\hat{\mathbf{Y}}|+|\mathbf{Y}|}$), which is also the F1 score in this context, Jaccard-index ($\frac{|\hat{\mathbf{Y}} \cap \mathbf{Y}|}{|\hat{\mathbf{Y}} \cup \mathbf{Y}|}$), Precision ($\frac{TP}{TP+FP}$), Recall ($\frac{TP}{TP+FN}$), Accuracy ($\frac{TP+TN}{TP+TN+FP+FN}$) and Matthews correlation coefficient ($\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$).

In this context, $TP$, $TN$, $FP$ and $FN$ represent pixels that are correctly classified as positive and negative, as well as pixels that are incorrectly classified as positive and negative, respectively. Further model evaluation was performed through constructions of ROC and Precision-Recall curves and through qualitative inspection of model outputs in cases where outputs closely resemble the label mask, and cases where outputs clearly differ from the label mask.

Finally, to assess the model's ability to identify TILs qualitative model output analysis was carried out. H&E patches of TIL regions were processed through the model, and the predictions of CD45RO-positive lymphocytes in proximity to epithelial tissue were analyzed by comparing the model output to the ground truth data. The decision to utilize CD45RO-positive lymphocytes for this evaluation aligns with the findings of Hu *et al.*, who demonstrated that tumor infiltration of CD45RO-positive lymphocytes is associated with favorable prognosis.

# 3 Results

The aim of this study was to evaluate the extent to which a U-NET model is able to predict the presence of various biomarkers utilizing masks derived from multiplexed fluorescence imaging.

## 3.1 Image registration

In order prepare data labels for training the segmentation model, image registration was performed through Warpy to align the CODEX multiplexed fluorescence data with H&E scans. For this study, we used two pairs of H&E scans and 7-plex CODEX data (table 1). Qualitative registration results for several patches are presented in Figure 6.

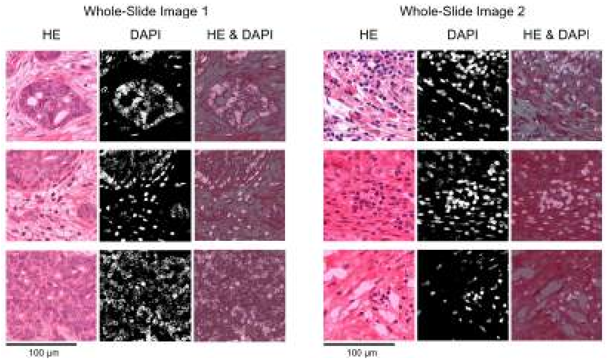

**Figure 6: Registration results of 3 patches per whole-slide image.** Rows present H&E, DAPI and a superimposition of H&E and DAPI for qualitative comparison. One summary is given for each of 2 whole-slide images used as data for this study. Scale bar is present indicating a span of 100 $\mu m$.

DAPI signals appear to overlap with nuclei in the associated H&E images, however pixel-precision inaccuracies are difficult to visually interpret. In order to quantify the improvement of the implemented image-registration workflow in comparison to traditional registration methods, haematoxylin stains were deconvolved out of H&E whole-slide images, thresholded through Otsu's method and registered onto a manually thresholded DAPI signal mask through rigid, affine and landmark-based thin-plate spline registration. The images were tiled into patches of 512 by 512 pixels and the Dice Similarity Coefficient was calculated for all registration methods per patch in order to measure similarity between the haematoxylin and DAPI channels. Distributions of these scores are presented as a violin-plot in figure 7 for relative similarity comparisons. Significance of differing distributions was determined through the Kolmogorov-Smirnov test.

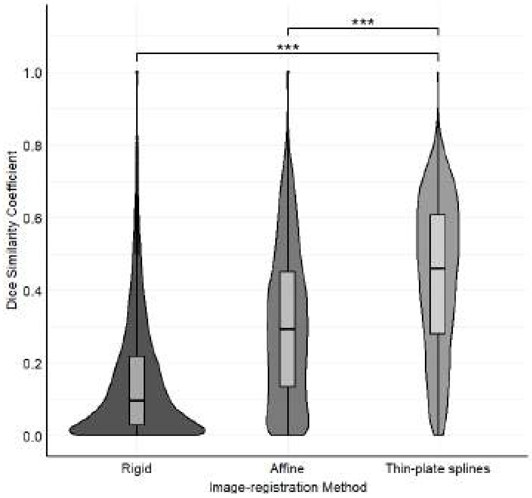

**Figure 7: Violin plot displaying distributions of different registration methods onto haematoxylin mask.** Image-registration methods are displayed on the x-axis. y-axis displays calculated Dice Similarity Coefficient. Box plots display inter-quartile and outlier ranges. Significance levels are displayed above the plots as calculated according to the Kolmogorov-Smirnov test ($p^{***} < 0.001$).

Distributions of Dice scores in patches registered through thin-plate splines are significantly higher than those of traditional registration methods, with a median score of 0.461 in contrast to median scores of 0.097 and 0.293 for rigid- and affine-transformations respectively. These findings suggest a significant enhancement in image-registration accuracy following the adoption of thin-plate spline landmark registration. For this reason, subsequent image-registration was performed through this method.

## 3.2 Whole-slide image patch-extraction

As part of the dataset preparation, patches were generated on whole-slide images through dlup(Teuwen et al. 2024), which employs the FESI-algorithm for foreground-background segmentation(Bug, Feuerhake, and Merhof 2015). In order to divide the dataset into manageable tiles for segmentation training, patches were created of 512 by 512 pixels with no overlap at a native resolution of $0.25 \mu m/px$. An overview of the tiles created through dlup are presented in figure 8.

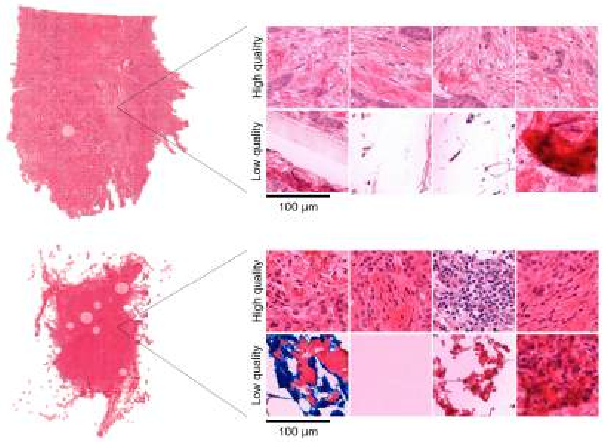

**Figure 8: Tiling overview of 2 whole-slide images used in this study.** Images are overlaid with red square patches indicating tiles. Each whole-slide image is paired with examples of high- and low quality patches, indicating artefacts present in the training data. Whole-slide images include empty circles remaining from previous TMA extraction. Scale bar is present indicating a span of 100 $\mu m$.

This figure displays the areas of the image that were extracted as patches in the dataset. Patch examples present the presence of artifacts such as staining residues, damaged and folded tissue areas, debris and empty regions from TMA-core extraction not excluded by the foreground segmentation algorithm. While dlup allows for mask thresholding, which increases stringency for areas being included in the algorithm, this may lead to loss of data in the final segmentation. For the purpose of this study, the dataset was carefully inspected for tiles including major artifacts and empty regions. Patches containing no meaningful information for the U-Net to train on were excluded from the dataset, while minor tissue damages and staining residues were included for potential generalizability of the model.

## 3.3 Label mask preparation

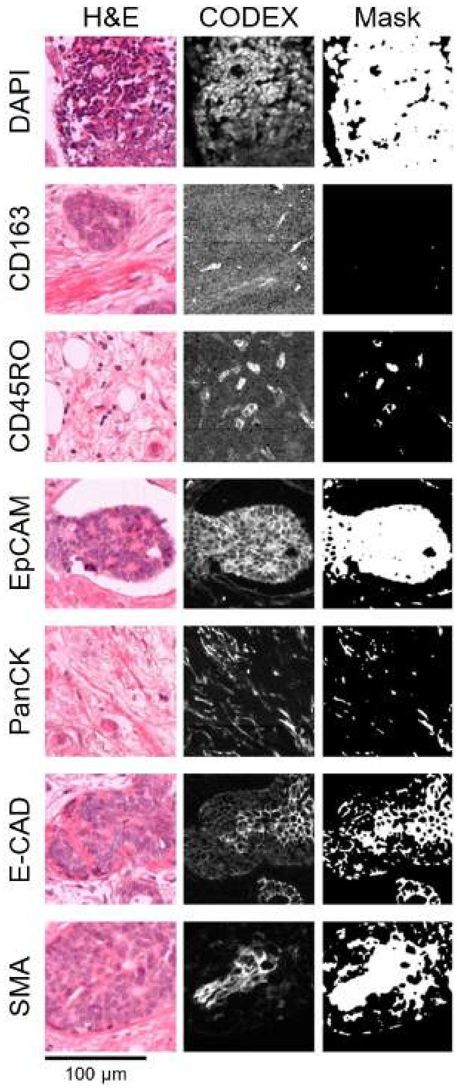

**Figure 9: General overview of masks displaying poor quality after applying global threshold.** H&E, normalized fluorescence and thresholded fluorescence masks are shown as one channel per row for all channels in the CODEX data. Mask image foreground is shown in white. Scale bar is displayed at the bottom of the figure indicating a span of $100 \mu m$.

In order to create ground-truth masks from the fluorescence CODEX data, a global threshold was applied to the image, followed by a 4x4 median filter to de-noise the image mask. This method worked well for label generation in most patches, however due to variation in image intensity across different regions in the slide image some mask regions fail to create a proper contrast between signal and background values. A range of poor-quality mask regions are shown for every channel in figure 9. In the displayed DAPI, EpCAM and SMA channel regions, the global threshold was too low and failed to divide the signal and the background. In the CD163, CD45RO and PanCK channels, the global threshold appears to be too high, and fails to capture signals present in the normalized CODEX fluorescence display. The E-CAD channel presents a challenge due to variability in signal intensity within a single structure, which cannot be extracted through a single threshold.

In order to create a general overview of the fluorescence intensity distribution and display the inconsistent intensity across the whole-slide images, each image was divided into tiles of 2048 by 2048 pixels. The median intensity of each tile was then calculated per channel and utilized to create an intensity map per channel for each image. An overview of the intensity distributions is presented in figure 10. These intensity maps illustrate a heterogeneous distribution of fluorescence intensity across the whole-slide images, indicating a root cause for inconsistent results through globally thresholded binary mask generation. While heterogeneous distributions are expected as a result of biomarker localisation, the distributions display signal aggregation around artefacts and outer edges of the tissue slide. Furthermore, these intensity maps reveal vertical signal-inconsistencies caused by the imaging process used to capture the data.

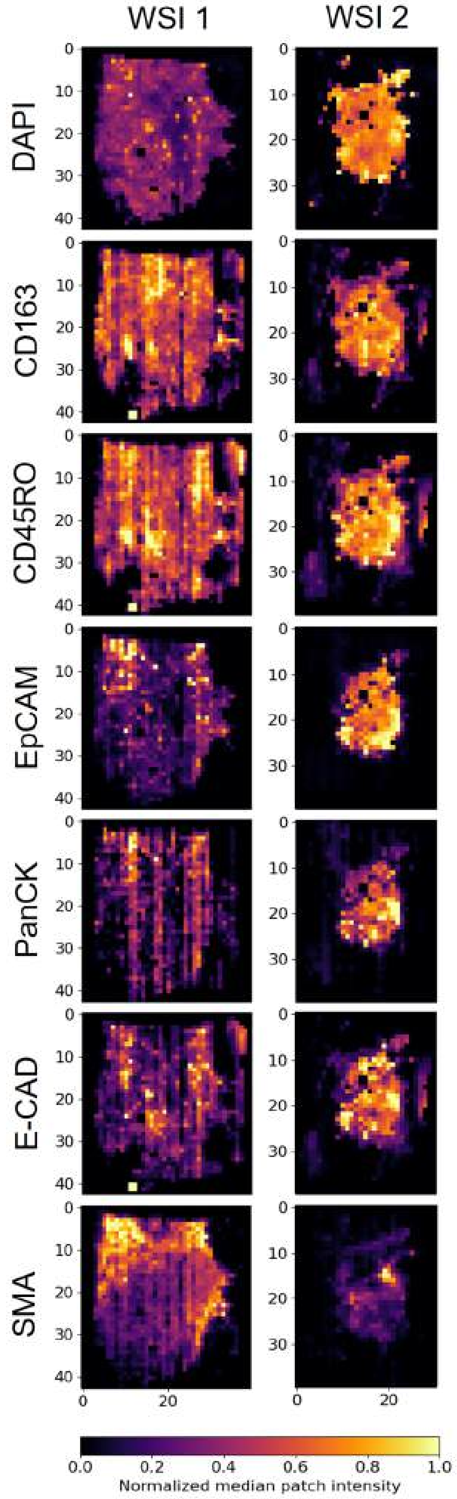

**Figure 10: Tiled intensity-maps of CODEX multiplexed fluorescence whole-slide images.** 2048 by 2048 pixel tiles showcasing normalized median fluorescence intensity after background subtraction and outlier-clipping. All 7 used fluorescence channels are displayed for both whole-slide images processed in this study. Tile indeces are presented on the x- and y-axis.

## 3.4 Model evaluation

To evaluate the capability of a U-Net segmentation model to predict the presence of various biomarkers, individual models were trained for each biomarker, with the exception of epithelial masks. Epithelial masks were derived from the combined presence of PanCK, E-CAD, and EpCAM biomarkers. Model performance metrics were calculated based on processed inference outputs, obtained by applying a sigmoid activation function to the model logits. Optimal threshold values for each model were determined by maximizing the $F_1$ scores on the validation set (figure s2). Each model output was then thresholded using its respective optimal threshold and evaluated based on a range of metrics as detailed in Table 2.

The Epithelial segmentation model demonstrates the highest performance with an $F_1$ Score of 0.769 and Jaccard Index of 0.629, indicating accurate segmentation of epithelial regions with balanced precision and recall (0.724 & 0.829). Similarly, the DAPI segmentation model shows strong performance, achieving an $F_1$ Score of 0.718 and Jaccard Index of 0.564, supported by high accuracy (0.912) and precision & recall (0.695, 0.748). In contrast, the CD163 segmentation model performs the poorest with a low $F_1$ Score of 0.035 and Jaccard Index of 0.018, despite high recall (0.960), reflecting significant challenges in precision (0.018) for CD163-positive regions. The CD45RO model exhibits similarly low performance, with an $F_1$ Score of 0.171 and Jaccard Index of 0.101, with an unbalanced precision and recall (0.109) & 0.612). The SMA model displays moderate performance with an $F_1$ Score of 0.353 and Jaccard Index of 0.217, highlighting the need for enhanced precision (0.252) while maintaining moderate recall (0.629).

In order to illustrate the model's performance further, ROC and Precision-Recall (PR) curves illustrating model predictions on their respective test sets are presented in figure 11. While DAPI and epithelium models present a sufficient AUC in both curves, the PR-curves show that CD163 and CD45RO models demonstrate a highly imbalanced recall and precision resulting in poor model performance. This imbalance of precision and recall suggests poor prediction of actual CD163+ and CD45RO+ signal.

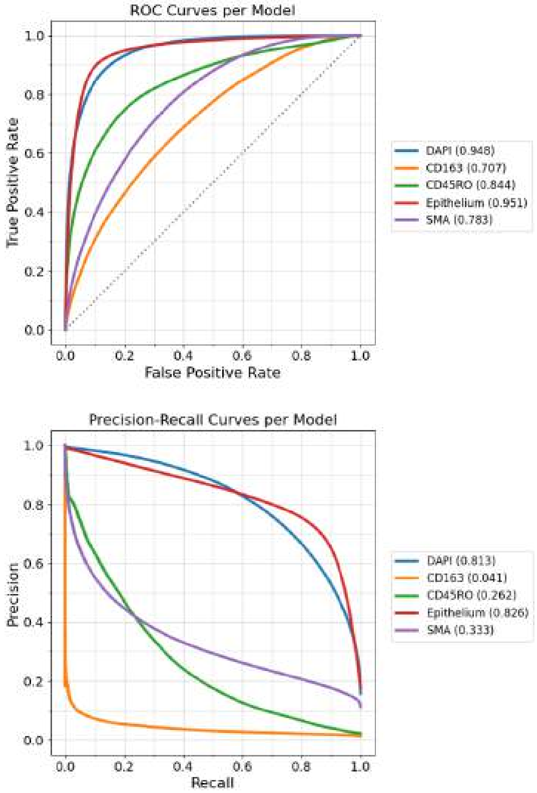

**Figure 11: ROC and Precision-Recall Curves for testing-set predictions of all trained models.** ROC curve x- and y-axes display false- and true-positive rate of the model predictions respectively. A hypothetical random classifier score is presented as a dotted line. Precision-Recall curve x- and y-axes display the recall and precision values for the model predictions respectively. A legend presents the respective models displayed with their AUC-values.

The SMA model displays a rapid decrease in its true-positive rate as the false-positive rate decreases. Similarly, it displays a strong increase in precision as recall decreases.

**Table 2: Evaluation metrics of models trained on each specific bio-marker.** Bio-markers are listed in the first column. A range of metrics are presented for each segmentation model. Rows are sorted in descending order by $F_1$-score.

| Bio-marker | $F_1$-Score | Jaccard-Index | Precision | Recall | Accuracy | MCC |
|---|---|---|---|---|---|---|
| Epithelial | 0.769 | 0.629 | 0.724 | 0.829 | 0.916 | 0.722 |
| DAPI | 0.718 | 0.564 | 0.695 | 0.748 | 0.912 | 0.667 |
| SMA | 0.353 | 0.217 | 0.252 | 0.629 | 0.740 | 0.270 |
| CD45RO | 0.171 | 0.101 | 0.109 | 0.612 | 0.866 | 0.206 |
| CD163 | 0.035 | 0.018 | 0.018 | 0.960 | 0.217 | 0.047 |

This behaviour indicates a possible bias towards the majority class, which in this case is the negative predictions in the segmentation. While outperforming the CD163 and CD45RO models, the SMA model likely suffers from an imbalance in its dataset.

## 3.5 Qualitative model analysis

In order to gain a better understanding of the model's predictions, several areas of a novel CODEX-paired whole-slide image were analysed qualitatively by comparing ground-truth image masks with similarly processed model outputs. This image was not included in the training data of the model, as the data was provided after training and evaluation were completed. In order to create these model outputs, novel 512 by 512 px H&E patches were processed through the U-Net models. Model outputs were thresholded using the previously determined optimal validation thresholds. Binary images from each channel were assigned specific colors and sequentially overlaid on top of one another in the following order: SMA, Epithelium, DAPI, CD45RO and CD163. An overview of model predictions for regions showing moderate, poor and artifact-specific model-predictions are presented in figure 12. In the following paragraph, images will be referred to as sub-images present in this figure.

Figure 12A displays moderately accurate model predictions, though a strong underestimation of CD163 and CD45RO is observed in these predictions. While images A-1, 2, 3 and 5 display clear CD163-positive signals, the CD163-model only predicts 2 small signals in A-1 within close proximity to the ground truth signals. Several small CD45RO-positive signals are found in the ground truth labels of images A-1, 2 and 3, but the CD45RO-model did not classify any pixels in these images as CD45RO-positive. Finally the epithelium- model is able to predict epithelium regions from the H&E patches consistently accurately, with the exception of image A-4 where the model displays an overestimation of the epithelium signal in this region.

Figure 12B presents H&E regions where the model performed poorly on matching ground-truth masks in its predictions. Image B-1 displays a region where the SMA- and epithelium-specific models produce overestimations of their respective signals. Here, SMA is particularly overestimated with the model classifying over 50% of the image as SMA-positive whereas the ground truth indicates the entire region to be negative for this signal. Interestingly, image B-2 shows the opposite outcome to be true where a SMA-positive signal is present but the model does not correctly identify it. Images B-2, 4, and 5 demonstrate the DAPI-model's inability to accurately identify the DAPI-positive regions within the shown H&E patches, highlighting challenges in extracting nuclear signals in these examples and vastly overestimating the positive region.

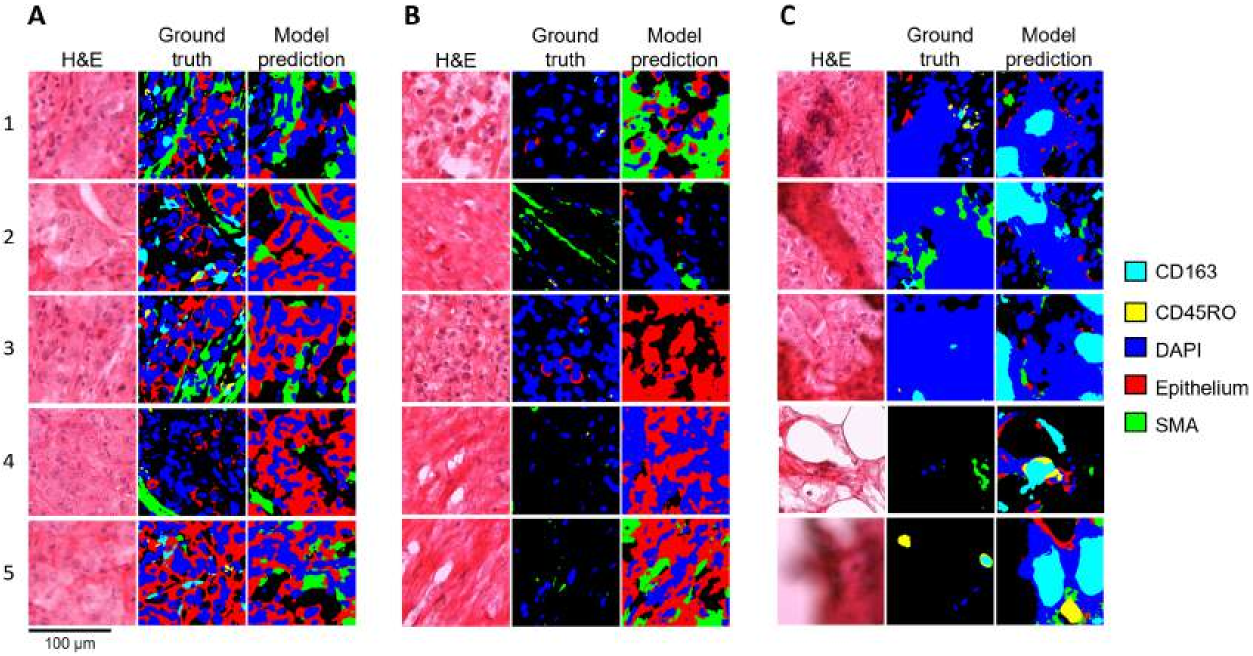

**Figure 12: Combined overview of moderate (A), poor (B) and image-artifact-specific (C) model predictions.** Each overview displays the input H&E image, CODEX-derived ground-truth labels and similarly processed model-outputs. Each row is labeled with an index for main-text referencing. Colours corresponding to individual bio-marker signals are referenced in the layered order in a legend on the right. Scale bar is displayed at the bottom of the figure indicating a span of $100\mu m$.

The opposite effect is illustrated in figure B-3, where nuclei are easily identifiable in the H&E region, but the model struggles to correctly correlate the nuclei to DAPI signal. Finally, images B-3, 4 and 5 display a substantial overestimation of epithelium in these H&E regions where ground truth masks indicate no epithelial-positive signal.

Figure 12C shows predicted bio-marker signals on image-artefacts present in the new whole-slide image. Given that a small number of image-artifacts were present in the training data, signal aggregation in damaged and folded tissue likely biased models towards positive-class prediction for these regions. Interestingly, the CD45RO and CD163 models made no substantial positive predictions on regular H&E regions but display a high confidence in predicting artifacts as regions positive for the respective bio-markers. Due to the large regions of DAPI-positive predictions covering most of the prediction in image C-1, 2 and 3, these images are displayed as individual channels in figure s3. Results show a positive prediction for markers SMA, Epithelium, DAPI and CD163 in images C- 1, 2 and 3 and for all 5 markers in images C-4 and 5 on image artifacts. These findings suggest there may in fact be a bias present in the models towards positive predictions on image artifacts caused by signal aggregation in these regions.

Finally, in order to assess the models' ability to identify tumour-infiltrating lymphocytes (TILs) co-localisation was analysed between CODEX-derived positive signals for Epithelial signal (EpCAM, PanCK, E-CAD) and CD45RO-positive cells. Several patches in these regions were qualitatively analysed in a comparative manner between CODEX-derived ground truth masks and processed model output. An overview of these comparisons is presented in figure 13.

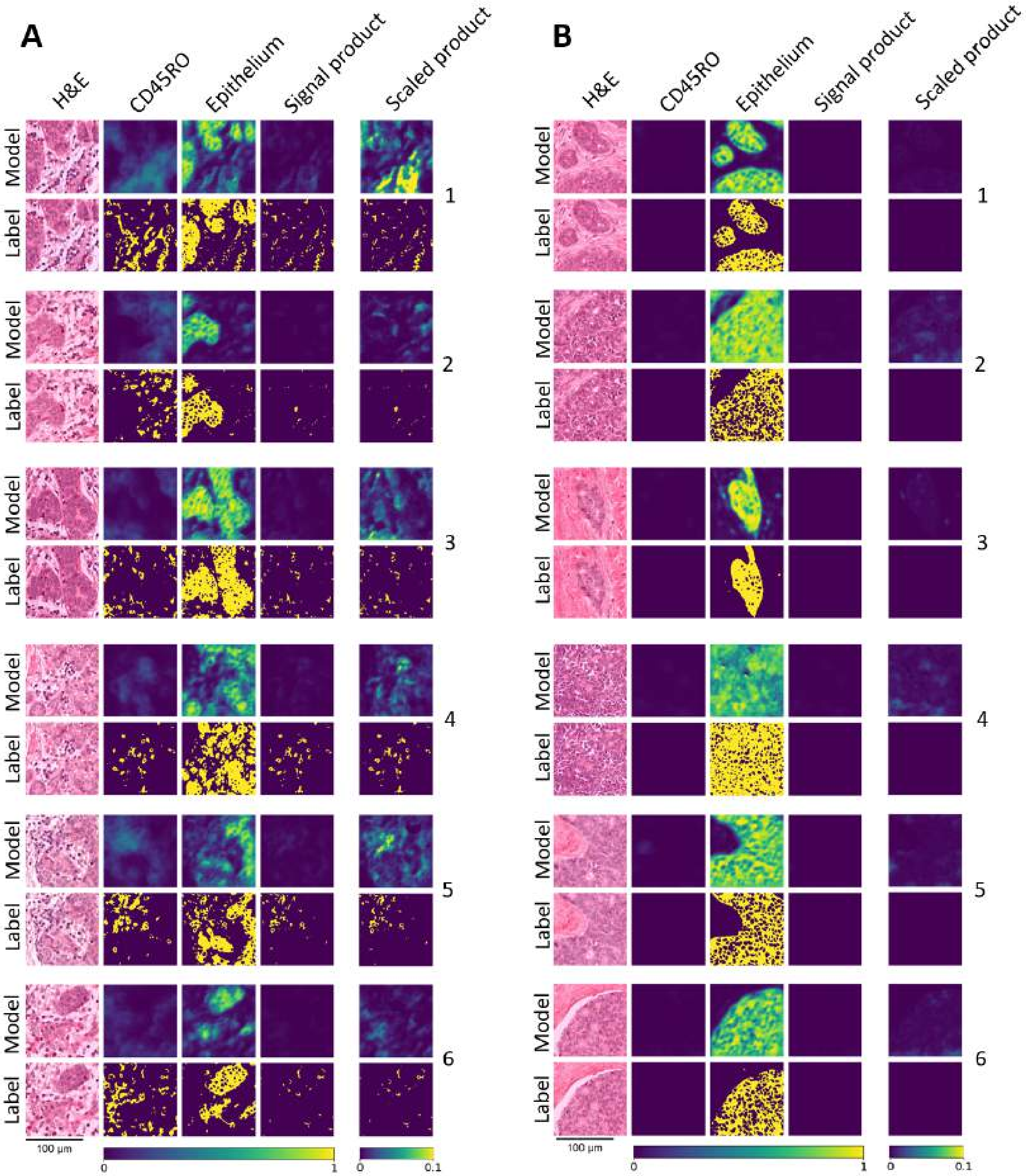

**Figure 13: Comparative overview of model output and ground truth labels for CD45RO$^{+}$ TIL-Positive regions (A) and TIL-Negative regions (B).** Each sub-figure displays six regions, with model outputs in the top row and ground truth labels in the bottom row. The columns, from left to right, present the H&E input, CD45RO signals, Epithelium signals, the element-wise product of the previous signals, and the same output normalized to a viewable max-range. Masks are displayed where 0 indicates background and 1 indicates signal. Model outputs are logits after a sigmoid activation function. Sub-figure indeces are present for main-text referencing. A scale bar at the bottom indicates a span of $100\mu m$.

From Figure 13 we generally observe higher confidence in CD45RO model outputs from epithelial regions containing CD45RO$^+$ TILs than regions containing epithelium with no lymphocyte infiltration. However, model outputs show an inability to pinpoint exact CD45RO-positive regions due to a low prediction confidence. As a result, we observe the same effect in the signal products. While figure 13A shows a general prediction of CD45RO localisation, it is not able to predict the exact signal of CD45RO label masks. Furthermore, Epithelium outputs generally show accurate prediction of epithelial label masks but often tend to overestimate the epithelium signal in empty regions slightly, which influences the predictions of lymphocyte and epithelium co-localisation. TIL region prediction in images A-2, 5 and 6 appear particularly overestimated due this false-positive epithelium prediction.

Figure 13B presents generally low CD45RO signal prediction in TIL-Negative regions, also leading to generally lower signal products. Slight signals in images B-2:6 are the effect of minor CD45RO-positive preditions being multiplied with epithelium predictions at high confidence. Scaled products in these figures display low-confidence TIL predictions overlapping with epithelium predictions. Generally, the model appears to distinguish between CD45RO-positive and negative regions, though confidence in its predictions is lacking, hindering proper co-localisation prediction.

# 4 Discussion

The primary goal of this study was to train a U-Net model on H&E-stained histopathological whole-slide images using CODEX-derived fluorescence as labeled data. We aimed to evaluate the extent to which the model could predict the presence of bio-markers from these generated labels and training data. To utilize these images as training data, we developed a comprehensive method involving patch extraction, image registration, label mask generation, and data filtering procedures.

## 4.1 Methodology challenges

From our employed patch-extraction method, we observed instances of patches containing image artifacts (figure 8) and noted the presence of aspecific fluorescence signals on artifacts (figure 12C). These results highlight the need for a better patch-extraction method, as manually filtering artifacts out of the dataset is unfeasible with increasing size of the dataset. Moreover, we found a significant improvement in image-registration performance after applying thin-plate spline deformations. Using Warpy and Elastix, we employed gradient-descent based optimization of the dice-score between haematoxylin and DAPI signals. This method outperformed rigid and affine transformations using an identical optimization algorithm, indicating that affine registration is not sufficient to register whole-slide images at cellular precision. To generate binary label masks from fluorescence data, we applied a global custom-thresholding method with a median-filter based post-processing step. From qualitative analyses, we observed several regions to be thresholded too high or too low, causing inconsistent mask labels (figures 9 and 10). These findings, along with observing inconsistent signal medians in vertical orientation along the codex-data, highlight the requirement for an improved mask-generation method focused on local fluorescence regions, as inconsistent data labels will significantly hinder model training.

## 4.2 Model performance

After utilizing the generated data for model training, several key observations were made. Models focusing on CD163, CD45RO, and SMA bio-markers showed difficulties balancing precision and recall and displayed the lowest AUC values in both the ROC and Precision-recall curves for all models (figure 11). Additionally, calculated metrics such as the $F_1$-Score, Jaccard-Index, and MCC were low for these models (Table 2), indicating poor performance. These findings are supported by qualitative analysis (figure 12 A and B), where SMA is often predicted in wrong regions (figure 12B), and CD163 and CD45RO show no predictions at all. Furthermore, all models display difficulties in correctly predicting any signal for the regions illustrated in figure 12B, possibly caused by darker red tones in elongated structures of the H&E patches. These colours vaguely mimic those found in the image artifacts in figure 12C, leading to inconsistent predictions caused by the positivity bias in artifacts. In contrast to the aforementioned under-prediction of CD163 and CD45RO, figure 12C shows large predicted regions of positivity for these bio-markers, suggesting that the model was influenced by the presence of image artifacts in the training data, leading to a bias towards these signals and a resulting under-prediction of true bio-marker signals.

Furthermore, we analyzed the ability of the models trained on epithelial and CD45RO signals to detect each respective bio-marker and examined their predicted and true co-localization (figure 13). We observed a general prediction of TIL localization, but the model did not predict these regions with high enough confidence to distinguish positive predictions in TIL-Positive regions from those in TIL-Negative regions. This is likely due to the CD45RO model's low confidence in its predictions, influenced by the image artifact-induced bias, and by slight over-prediction of epithelium-negative areas by the epithelium model.

## 4.3 Limitations

These findings come with a number of key-limitations: Firstly, while our image registration method shows improved results, it assumes that the calculated transformation of the DAPI channel will yield accurate registration for every channel in the data, while this may not always be true. Slight errors in pixel-precision image registration may break the signal between H&E data and the generated label masks, suggesting this method may be further improved. Furthermore, only 2 whole-slide images were used in training the models for this study and no TMA cores were available to be utilized as training data. This has likely led to over-fitting of the models on only these 2 images, or ideally, to only slide-images of tumours with these visual characteristics. Additionally, due to a lack of data quantity, training, validation and test-set data were all retrieved from identical whole-slide images. While we ensured there would be no overlap between training and testing patches, tiles from different sets will have similar tumour-characteristics. This has likely led to overconfident model evaluation, as no true generalization has been shown during model evaluation apart from qualitative inference analysis in figure 12. This is because this analysis was performed on a whole-slide image and codex-data pair which was provided after model development and evaluation was completed. Finally, as this newly provided whole-slide image did not display lymphocyte infiltration of epithelial regions, TIL analysis was performed on one of the whole-slide images the model was trained on. This leads to overestimation of the actual performance of the model in predicting TIL-localisation, concluding that TIL-analysis results are not reliable and should be studied on a novel dataset instead.

## 4.4 Future directions

In order to address potential misalignment of the fluorescence channels, raw CODEX data is paired with a DAPI signal for every cycle of codex antibody binding. For optimal registration, one Warpy-based transformation file may be created per CODEX cycle, and the CODEX data can be registered in groups of 4-channels per transformation. One drawback introduced by this method is an increase in computation time and manual data processing, as Warpy registration requires human input in order to perform a proper transformation. Furthermore, exporting of multi-channel whole slide image data is time consuming, and dividing this export into multiple steps creates an increase in export time and method complexity. One consideration to address this is to register the data through the first and last cycle DAPI measurement, and linearly adjusting for sample drift between these cycles. This method would eliminate the need to perform many registrations, but would not eliminate the increased export time.

Exploring prior research into image artifacts reveals artifacts in histopathological whole-slide images is a common issue in the field, and multiple efforts are currently being made to detect these artifacts for future filtering or processing (Smit et al. 2021; Kanwal et al. 2024). Detecting image artifacts in order to ensure quality of the extracted patches in this method, as well as increasing the background-mask threshold to limit edges of the slide image to be included in the dataset, are important for ensuring proper quality of the data prior to training segmentation models. We have demonstrated that models exhibit bias towards these artifact regions, displaying high prediction confidence on image artifacts for multiple channels. Improving the patch extraction in our method may increase model performance and decrease artifact-induced biases. Schömig *et al.* (Schömig-Markiefka et al. 2021) demonstrates a decrease in model performance influenced by the presence of many types of artifacts present in training datasets. They explicitly state that most deep-learning methods do not control for artifact effects, and that this leads to bias through selection of higher-quality slides. This statement highlights the need for generalized high performance on any selection of tissue slides in order to deploy models in routine classifications. While the method described in this study may be simply improved by including only whole-slide images of high quality, it may be more favourable for model generalisation to employ a standard artifact detection algorithm or a deep-learning network trained to detect artifacts in whole-slide images to filter these artifacts out of the dataset.

To explore mask generation techniques, multiple papers in the field of image-processing employ k-means clustering (MacQueen et al. 1967) in order to segment differing regions within one image (Sarrafzadeh and Dehnavi 2015; McColloch et al. 2019). Employing k-means clustering may improve quality of masks where fluorescence shows inconsistency, as shown in figure s4, however this presents a few problems. Firstly, up-scaling mask generation to entire whole-slide images may be unfeasible as clustering algorithms are computationally intensive on large-scale data, furthermore determining the optimal number of clusters to yield correct masks is not always reliably automated. Literature on whole-slide image fluorescence masking is scarce, and no workflows exist for this purpose to our knowledge. In order to accurately generate image masks per whole-slide image region, Pyvips (Martinez and Cupitt 2005) could be employed to extract smaller image regions and pre- and post-processing these regions before and after either applying a locally calculated threshold or using k-means clustering using an optimal k-value determined through a reliable metric. Processed masks may then be iteratively written through Pyvips.

Finally, while class imbalance was handled through under-sampling of the majority negative class by excluding many empty mask labels from the datasets, future efforts to address the imbalance may include a weighted loss function to penalize misclassifications of the minority class more heavily. This approach aims to ensure that the model learns to better predict underrepresented bio-markers, thereby improving overall performance on the bio-markers exhibiting low performance.

## 4.5 Summary

In summary, these findings indicate that while the method developed for this study shows promise, significant challenges remain. The presence of image artifacts, inconsistent label generation, and limited data quantity and quality all contribute to model performance issues particularly in underrepresented bio-markers such as CD45RO and CD163. Future research should focus on improving patch extraction, label mask generation methods and handling class imbalance as well as increasing the amount of training data in order to improve model performance and generalization. As task complexity increases, additional model complexity may be considered, such as implementation of attention gates or increasing U-Net depth. Finally, utilizing separate training, validation and testing whole-slide images is crucial to obtain more accurate and unbiased model performance metrics.

While these preliminary results suggest broad room for improvement, our findings have demonstrated several instances of moderately accurate prediction of DAPI, epithelium, and smooth muscle actin presence on data from a whole-slide image that the model was not trained on. With a range of improvements in this developed method to enhance data quality and an increase in training data quantity, this model may achieve performance levels accurate enough to assist trained pathologists in the assessment of tumor biopsies by providing precise cell classifications. Currently, pathologists perform the prognosis largely on their own, classifying different cell types and determining the potential presence of TILs (tumor-infiltrating lymphocytes). Implementing automated classification could significantly speed up prognoses for patients by having the initial classification done automatically, thus shifting pathologists to a quality control role. This highlights the necessity for expanding our current dataset and refining our proposed method. The evidence presented in this study supports the feasibility of these improvements, though future work will be needed to overcome practical challenges and validate these enhancements in a clinical setting.

# 5 Conclusion

This study aimed to develop and evaluate a methodology for training a U-Net model on H&E-stained histopathological whole-slide images using CODEX-derived fluorescence as labeled data. Our findings highlight several challenges and potential directions for improvement in the developed methodology.

Firstly, our approach revealed significant issues with image artifacts during patch extraction, leading to inconsistent labeling and subsequent bias in model predictions. Addressing this challenge requires improving our current patch-extraction method to better filter out artifacts automatically, thus improving data quality for model training. Furthermore, while our image-registration technique showed promising improvements with thin-plate spline deformations, ensuring accurate alignment across all fluorescence channels remains critical. Future data processing pipelines should ensure that registration is precise over all channels in the data. Additionally, the method for generating binary label masks from fluorescence data exhibited inconsistencies, particularly through global thresholding accuracy. Integrating advanced image-processing techniques such as k-means clustering or leveraging tools like Pyvips for local thresholding could enhance mask quality and improve model performance. Finally, evaluation of our models highlighted specific challenges in predicting underrepresented biomarkers like CD163 and CD45RO, where precision and recall remain unbalanced. Addressing class imbalance through weighted loss functions and expanding our dataset to include a wider variety of tissue samples are critical steps towards improving model robustness and generalization.

Our current model also demonstrates potential in identifying tumor-infiltrating lymphocytes, although our findings are somewhat overestimated due to the analysis being conducted on a whole-slide image used for model training. Optimizing our approach to include separate whole-slide images for training, validation, and testing will be essential to obtain accurate and unbiased assessments of TIL identification capabilities and general model performance metrics.

In conclusion, while our study demonstrates promise in automated biomarker prediction and TIL identification from histopathological images, substantial improvements are necessary to achieve reliable clinical applicability. Future research efforts should focus on refining data preprocessing steps, increasing dataset diversity and enhancing model robustness through advanced architectural modifications and different hyper parameters. These enhancements will be crucial in developing a reliable automated tool that can assist pathologists in tumor assessment and prognosis effectively.

# A Acknowledgements

All CODEX and H&E data acquisition for this study was performed by the Core Facility Molecular Pathology & Biobanking at Nederlands Kanker Instituut (NKI).

General assistance and computing power for model training were provided by the Bioimaging facility at NKI

Special thanks to the computational pathology group at NKI for assistance and support throughout this project

# B Availability

## B.1 Code repositories

General scripts used in this study are available at:

- https://github.com/BioImaging-NKI/fusion2ims
- https://github.com/Idrismania/image_processing_scripts

Code used for model training in this study, as well as model weights are available at:

- https://github.com/Idrismania/bioinformatics_project

## B.2 Methods scripts

`1: fusion_to_ims.py`
Python script for generating raw CODEX data from intermediate data files

`2: 01_export_HE.groovy`
Groovy script for exporting H&E images in QuPath

`3: 02_extract_8bit_DAPI.groovy`
Groovy script for exporting a single channel from a CODEX multi-channel image and converting to 8-bit in QuPath.

`4: 03_export_overlay_channels.groovy`
Groovy script for exporting a custom selection of channels from a concatenated image server in QuPath

`5: 04_export_masks_with_custom_thresholds.groovy`
Groovy script for exporting image masks through a series of custom threshold values in QuPath

`6: dlup_tiler.py`
Python script for pre-tiling whole slide images and corresponding masks

## B.3 Data availability

Data used for this project may be requested from Rolf Harkes at:

- r.harkes@nki.nl

# C References

Ahmedt-Aristizabal, David et al. (2022). “A survey on graph-based deep learning for computational histopathology”. In: *Computerized Medical Imaging and Graphics* 95, p. 102027.

Bankhead, Peter et al. (2017). “QuPath: Open source software for digital pathology image analysis”. In: *Scientific reports* 7.1, pp. 1–7.

Bhandary, Shrajan et al. (2023). “Investigation and benchmarking of U-Nets on prostate segmentation tasks”. In: *Computerized Medical Imaging and Graphics*, p. 102241.

Black, Sarah et al. (2021). “CODEX multiplexed tissue imaging with DNA-conjugated antibodies”. In: *Nature protocols* 16.8, pp. 3802–3835.

Bradski, G. (2000). “The OpenCV Library”. In: *Dr. Dobb’s Journal of Software Tools*.

Broeders, Mireille et al. (2012). “The impact of mammographic screening on breast cancer mortality in Europe: a review of observational studies”. In: *Journal of medical screening* 19.1_suppl, pp. 14–25.

Bug, Daniel, Friedrich Feuerhake, and Dorit Merhof (2015). “Foreground extraction for histopathological whole slide imaging”. In: *Bildverarbeitung für die Medizin 2015: Algorithmen-Systeme-Anwendungen. Proceedings des Workshops vom 15. bis 17. März 2015 in Lübeck.* Springer, pp. 419–424.

Bulten, Wouter et al. (2019). “Epithelium segmentation using deep learning in H&E-stained prostate specimens with immunohistochemistry as reference standard”. In: *Scientific reports* 9.1, p. 864.

Chan, John KC (2014). “The wonderful colors of the hematoxylin–eosin stain in diagnostic surgical pathology”. In: *International journal of surgical pathology* 22.1, pp. 12–32.

Chiaruttini, Nicolas et al. (2022). “An open-source whole slide image registration workflow at cellular precision using Fiji, QuPath and Elastix”. In: *Frontiers in Computer Science* 3, p. 780026.

Clemente, Claudio G et al. (1996). “Prognostic value of tumor infiltrating lymphocytes in the vertical growth phase of primary cutaneous melanoma”. In: *Cancer: Interdisciplinary International Journal of the American Cancer Society* 77.7, pp. 1303–1310.

Couture, Heather D (2022). “Deep learning-based prediction of molecular tumor biomarkers from H&E: a practical review”. In: *Journal of Personalized Medicine* 12.12, p. 2022.

Cui, Miao and David Y Zhang (2021). “Artificial intelligence and computational pathology”. In: *Laboratory Investigation* 101.4, pp. 412–422.

Echle, Amelie et al. (2021). “Deep learning in cancer pathology: a new generation of clinical biomarkers”. In: *British journal of cancer* 124.4, pp. 686–696.

Ferlay, Jacques et al. (2018). “Global cancer observatory: cancer today”. In: *Lyon, France: international agency for research on cancer* 3.20, p. 2019.

Goode, Adam et al. (2013). “OpenSlide: A vendor-neutral software foundation for digital pathology”. In: *Journal of pathology informatics* 4.1, p. 27.

Gooden, Marloes JM et al. (2011). “The prognostic influence of tumour-infiltrating lymphocytes in cancer: a systematic review with meta-analysis”. In: *British journal of cancer* 105.1, pp. 93–103.

Hendry, Shona et al. (2017). “Assessing tumor-infiltrating lymphocytes in solid tumors: a practical review for pathologists and proposal for a standardized method from the International Immuno-Oncology Biomarkers Working Group: Part 2: TILs in melanoma, gastrointestinal tract carcinomas, non–small cell lung carcinoma and mesothelioma, endometrial and ovarian carcinomas, squamous cell carcinoma of the head and neck, genitourinary carcinomas, and primary brain tumors”. In: *Advances in anatomic pathology* 24.6, pp. 311–335.

Hu, Guoming and Shimin Wang (2017). “Tumor-infiltrating CD45RO+ memory T lymphocytes predict favorable clinical outcome in solid tumors”. In: *Scientific reports* 7.1, p. 10376.

Kanwal, Neel et al. (2024). “Equipping Computational Pathology Systems with Artifact Processing Pipelines: A Showcase for Computation and Performance Trade-offs”. In: *medRxiv*, pp. 2024–03.

Kingma, Diederik P and Jimmy Ba (2014). “Adam: A method for stochastic optimization”. In: *arXiv preprint arXiv:1412.6980*.

Klein, Stefan et al. (2009). “Elastix: a toolbox for intensity-based medical image registration”. In: *IEEE transactions on medical imaging* 29.1, pp. 196–205.

Leong, Anthony SY, F Joel, and W-M Leong (2005). “Strategies for laboratory cost containment and for pathologist shortage: centralised pathology laboratories with microwave-stimulated histoprocessing and telepathology”. In: *Pathology* 37.1, pp. 5–9.

Lu, Ming Y et al. (2023). “Towards a visual-language foundation model for computational pathology”. In: *arXiv preprint arXiv:2307.12914*.

Łukasiewicz, Sergiusz et al. (2021). “Breast cancer—epidemiology, risk factors, classification, prognostic markers, and current treatment strategies—an updated review”. In: *Cancers* 13.17, p. 4287.

MacQueen, James et al. (1967). “Some methods for classification and analysis of multivariate observations”. In: *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*. Vol. 1. Oakland, CA, USA, pp. 281–297.

Martinez, Kirk and John Cupitt (2005). “VIPS-a highly tuned image processing software architecture”. In: *IEEE International Conference on Image Processing 2005.* Vol. 2. IEEE, pp. II–574.

McColloch, Andrew et al. (2019). “Correlation between nuclear morphology and adipogenic differentiation: application of a combined experimental and computational modeling approach”. In: *Scientific Reports* 9.1, p. 16381.

Mortazi, Aliasghar et al. (2023). “Selecting the best optimizers for deep learning–based medical image segmentation”. In: *Frontiers in Radiology* 3.

Nickolls, John et al. (2008). “Scalable parallel programming with cuda: Is cuda the parallel programming model that application developers have been waiting for?” In: *Queue* 6.2, pp. 40–53.

Oktay, Ozan et al. (2018). “Attention u-net: Learning where to look for the pancreas”. In: *arXiv preprint arXiv:1804.03999.*

Pagès, Franck et al. (2009). “In situ cytotoxic and memory T cells predict outcome in patients with early-stage colorectal cancer”. In: *Journal of clinical oncology* 27.35, pp. 5944–5951.

Paszke, Adam et al. (2019). “Pytorch: An imperative style, high-performance deep learning library”. In: *Advances in neural information processing systems* 32.

Pietzsch, Tobias et al. (2015). “BigDataViewer: visualization and processing for large image data sets”. In: *Nature methods* 12.6, pp. 481–483.

Ronneberger, Olaf, Philipp Fischer, and Thomas Brox (2015). “U-net: Convolutional networks for biomedical image segmentation”. In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18.* Springer, pp. 234–241.

Saltz, Joel et al. (2018). “Spatial organization and molecular correlation of tumor-infiltrating lymphocytes using deep learning on pathology images”. In: *Cell reports* 23.1, pp. 181–193.

Sarrafzadeh, Omid and Alireza Mehri Dehnavi (2015). “Nucleus and cytoplasm segmentation in microscopic images using K-means clustering and region growing”. In: *Advanced biomedical research* 4.1, p. 174.

Schindelin, Johannes et al. (2012). “Fiji: an open-source platform for biological-image analysis”. In: *Nature methods* 9.7, pp. 676–682.

Schömig-Markiefka, Birgid et al. (2021). “Quality control stress test for deep learning-based diagnostic model in digital pathology”. In: *Modern Pathology* 34.12, pp. 2098–2108.

Schwen, Lars Ole et al. (2023). “Digitization of pathology labs: a review of lessons learned”. In: *Laboratory Investigation*, p. 100244.

Smit, Gijs et al. (2021). “Quality control of whole-slide images through multi-class semantic segmentation of artifacts”. In: *Medical Imaging with Deep Learning*.

Sobral-Leite, Marcelo et al. (2019). “Cancer-immune interactions in ER-positive breast cancers: PI3K pathway alterations and tumor-infiltrating lymphocytes”. In: *Breast Cancer Research* 21, pp. 1–12.

Song, Andrew H et al. (2023). “Artificial intelligence for digital and computational pathology”. In: *Nature Reviews Bioengineering* 1.12, pp. 930–949.

Tellez, David et al. (2018). “Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks”. In: *IEEE transactions on medical imaging* 37.9, pp. 2126–2136.

Teuwen, J. et al. (May 2024). *DLUP: Deep Learning Utilities for Pathology*. Version 0.3.38. URL: `https://github.com/NKI-AI/dlup`.

Underwood, JC (1974). “Lymphoreticular infiltration in human tumours: prognostic and biological implications: a review.” In: *British journal of cancer* 30.6, p. 538.

Vaswani, Ashish et al. (2017). “Attention is all you need”. In: *Advances in neural information processing systems* 30.

# D Appendix

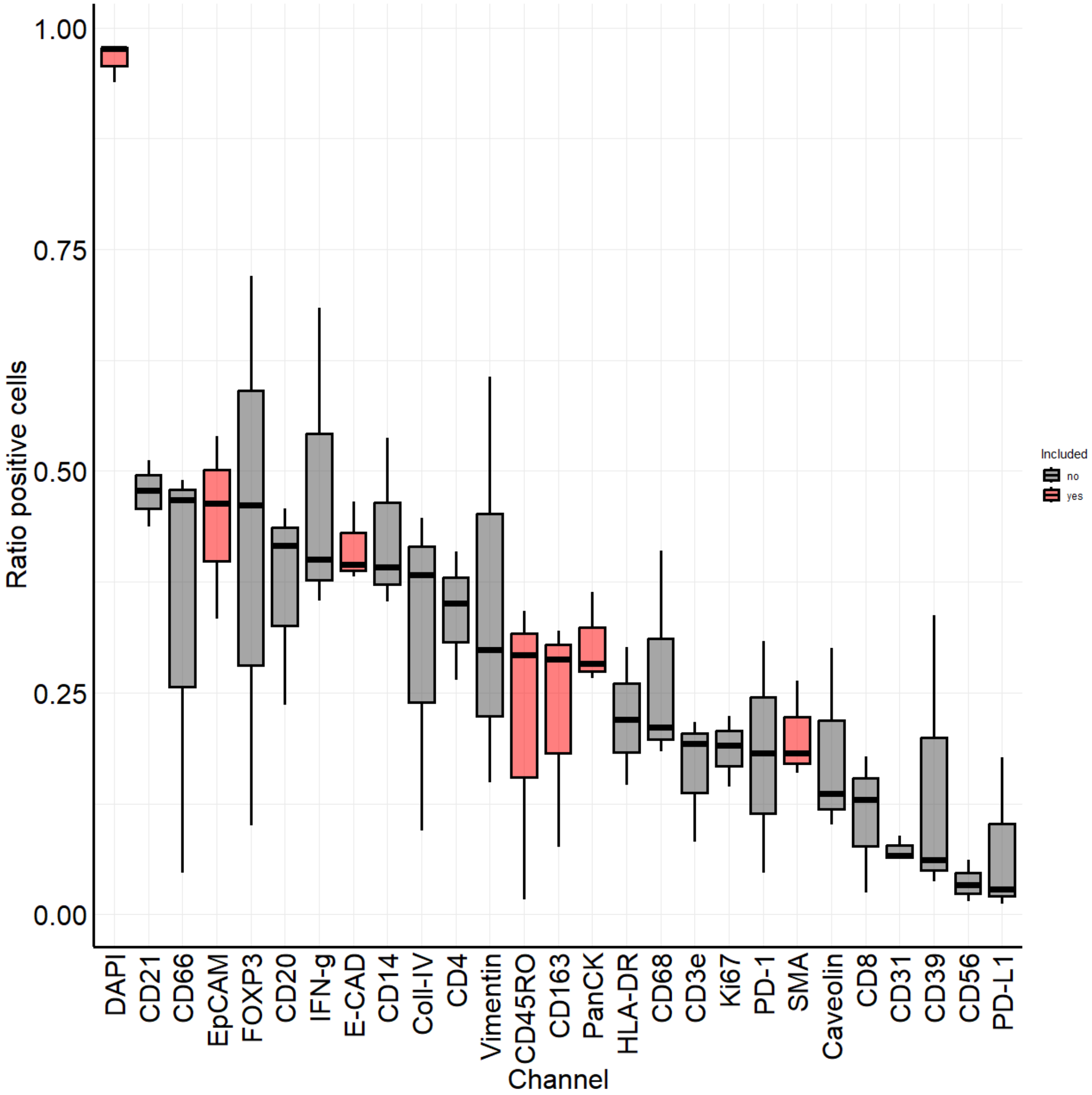

**Figure s1: Ratio of positive cells over multiplexed whole-slide images.** Channels are shown sorted by median in descending order on the x-axis. Box-plots present inter-quartile and outlier ranges of positive cell ratios per channel along the y-axis. Channels included in this study as a channel for model training are shown in red.

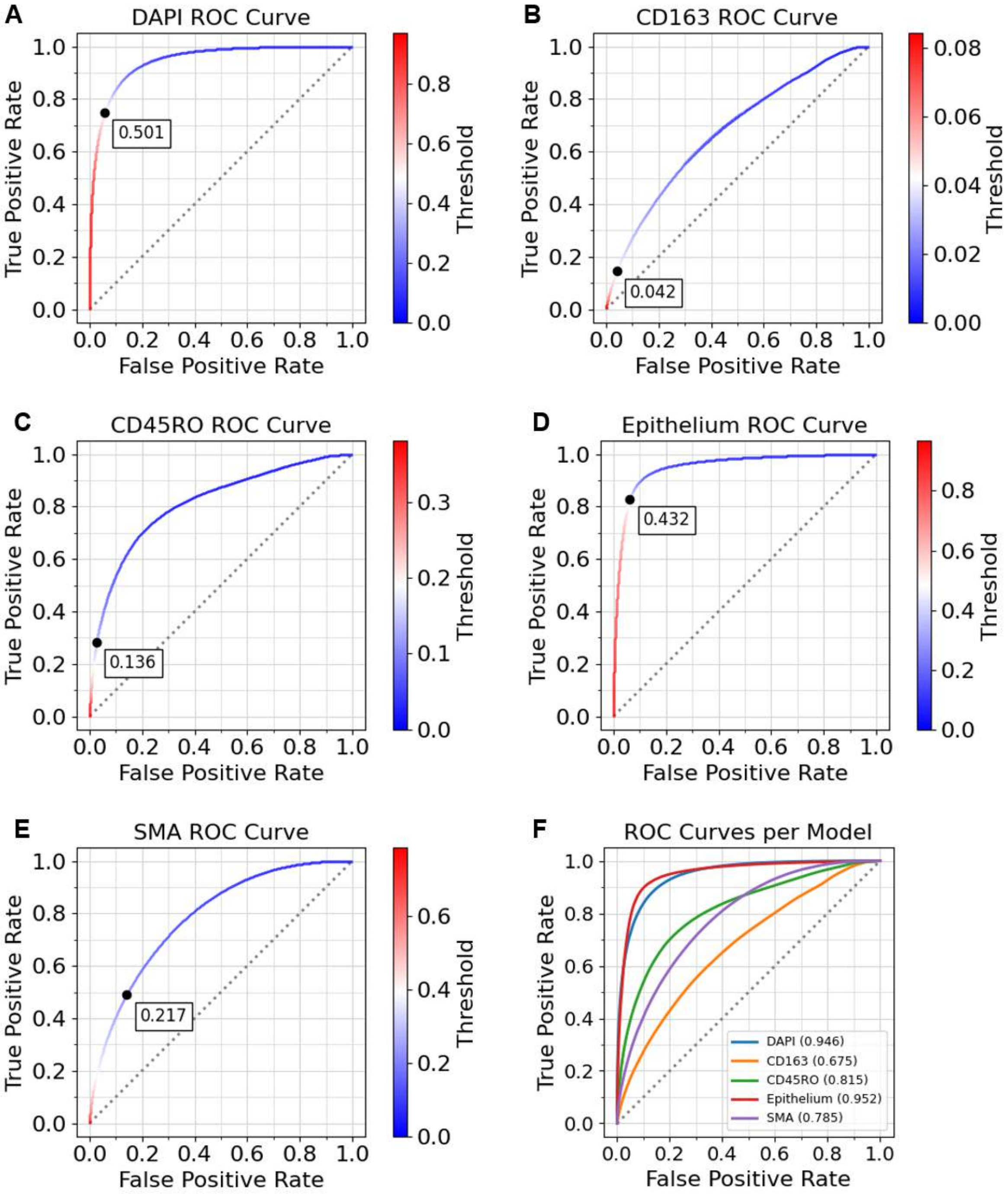

**Figure s2: Receiver operating characteristic (ROC) curve for validation-set predictions of all models trained in this study and a combined plot with AUC values reported.** x- and y- axes display false- and true-positive rate respectively. Plots A through E display respective threshold values along the curve. Color bar displays threshold range for each plot and optimal threshold for $F_1$-maximization is denoted with a point and value-annotation. A hypothetical random classifying score is plotted as a doted line. Plot F displays a combined plot of all ROC curves. Displayed legend presents AUC values for every curve plotted.

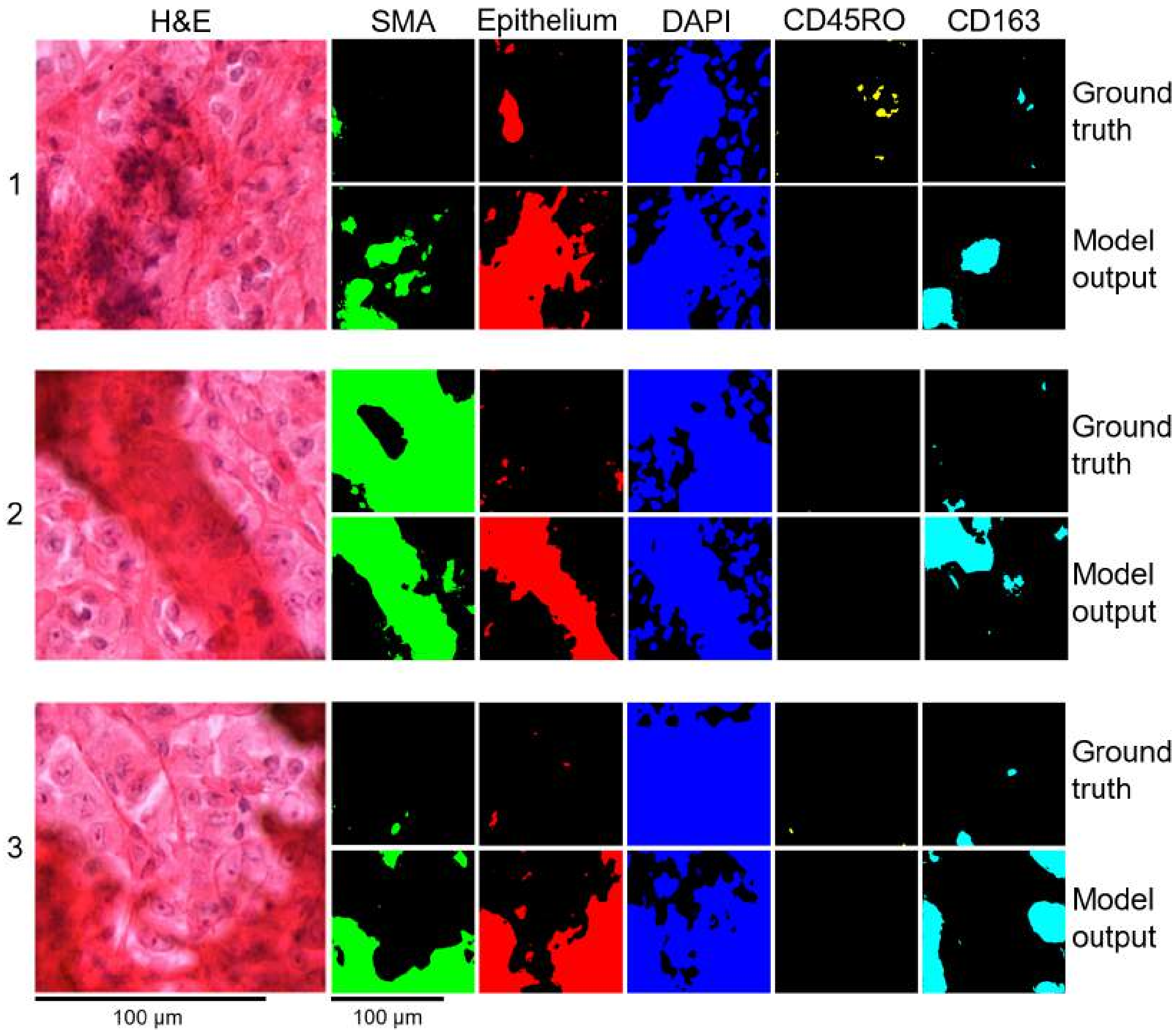

**Figure s3: Individual channel-displays for Figure 12C-1, 2 and 3.** Each channel is presented in the respective colour from the original figure. Columns represent different channels, with H&E presented on the left with 2x up-sizing. Each pair of rows displays the ground truth and the model output respectively. Image identifiers from original figure are displayed on the left. Scale bars are displayed at the bottom of the figure, each indicating a span of $100\mu m$.

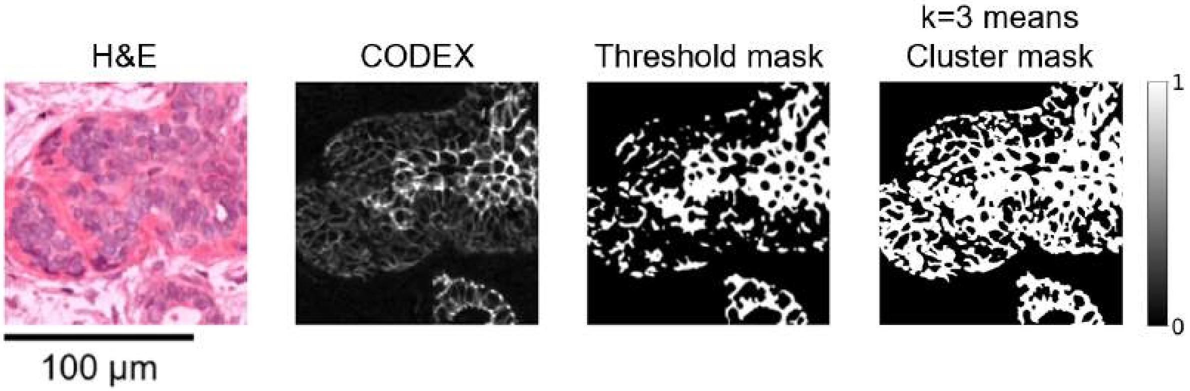

**Figure s4: Qualitative mask generation performance after applying k-means clustering on figure 9 E-CAD channel.** Image displays H&E region accompanied by corresponding CODEX-fluorescence, thresholded mask and k-means mask. Intensity ranges are illustrated by a color bar on the right. Scale bar is displayed at the bottom of the figure indicating a span of $100\mu m$.