Master's thesis
Master's Programme in Life Science Informatics

Mapping Regions of Interest in Ovarian Cancer: A Deep Learning Approach with Multiplexed Imaging

Matias Aiskovich

February 14, 2025

Supervisor(s): Anniina Färkkilä, MD PhD

University of Helsinki
Faculty of Science

P. O. Box 68 (Pietari Kalmin katu 5)
00014 University of Helsinki

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty Faculty of Science	Koulutusohjelma — Utbildningsprogram — Degree programme Master’s Programme in Life Science Informatics

Tekijä — Författare — Author

Matias Aiskovich

Työn nimi — Arbetets titel — Title

Mapping Regions of Interest in Ovarian Cancer: A Deep Learning Approach with Multiplexed Imaging

Työn laji — Arbetets art — Level Master’s thesis	Aika — Datum — Month and year February 14, 2025	Sivumäärä — Sidantal — Number of pages 59

Tiivistelmä — Referat — Abstract

Advancements in machine learning can transform digital pathology by enabling the automated identification of biologically significant regions in complex tissue images. This study presents a comprehensive benchmarking of deep learning models to uncover spatial histopathological and molecular patterns linked to chemotherapy exposure in high-grade serous ovarian carcinoma (HGSC).

A set of machine learning classifiers was trained on multiplex immunofluorescence (t-CycIF) images to determine whether tissue samples had been exposed to chemotherapy. Multiple Instance Learning (MIL) approach with pre-trained pathology encoders resulted in the best classifier performance. The model identified the high-confidence tumor regions most relevant for classification.

The results demonstrate that AI-driven region selection can be a key function of machine learning models in pathology, providing meaningful biological insights. This approach could enhance biomarker discovery, aid in patient stratification, and support the development of precision oncology strategies.

ACM Computing Classification System (CCS):
General and reference → Document types → Surveys and overviews
Applied computing → Life and medical sciences → Computational biology → Imaging

Avainsanat — Nyckelord — Keywords

HGSOC, multiplexed imaging, tumor microenvironment, bioinformatics, machine-learning

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — Övriga uppgifter — Additional information

Contents

1 Introduction **1**

1.1 Rationale . . . 2

1.2 Aim . . . 2

2 Background **5**

2.1 High Grade Serous Ovarian Carcinoma . . . 5

2.2 Primary Debulking Surgery and Interval Debulking Surgery . . . 5

2.3 Tumor microenvironment . . . 6

2.4 Biomarker-based therapeutics . . . 6

2.5 Tissue microarray . . . 7

2.6 Tissue Cyclic Immunofluorescence Imaging . . . 7

2.7 Computer Vision . . . 8

2.8 Machine learning for digital pathology . . . 9

2.9 Biological interpretation of ML features . . . 10

2.10 Variational Auto Encoders . . . 11

2.11 ResNet classifier . . . 12

2.12 Foundation models' (for pathology) . . . 12

2.13 Attention . . . 13

2.14 Multiple-instance learning . . . 13

2.15 F1-score . . . 13

3 Methods **19**

3.1 Computational Analysis Pipeline . . . 19

3.2 Data . . . 19

3.2.1 NKI dataset . . . 20

3.3 Preprocessing: Normalization of t-CycIF data . . . 20

3.4 Cohort selection . . . 22

3.5 Preprocessing: Tiling . . . 22

3.6 Classification task . . . 23

3.7 Model Comparisson . . . 24

3.8 Attention weights 25

4 Results 27

4.1 Tiling 27

4.2 Variational Auto Encoders 27

4.3 ResNet classifier 29

4.4 MIL classifier 31

4.5 Identified ROIs 32

5 Discussion 35

6 Conclusions 37

6.0.1 Future work 37

Bibliography 41

Appendix A NKI Antibodies list 49

Appendix B Models experiments 51

Appendix C Identified ROIs 53

Appendix D Signal variation in t-CycIF data 57

1. Introduction

Current technologies such as Cyclic Immunofluorescence images (t-CycIF) [29] allow for highly multiplexed imaging and measurement of protein levels in single cells. t-CycIF can produce images with up to 60 channels by sequentially staining, imaging, and then inactivating fluorescent dyes on the antibodies. This allows for the simultaneous measurement of many proteins in the same sample, with the resulting images with pixel-level data on protein quantities.

When working with ovarian cancer samples whole slide images (WSI), it can be expected to obtain images of more than 50000 × 40000 pixels on average [50]. If, with t-CycIF, 60 antibodies (channels) are run in these images, the resulting images will contain more than 100 billion pixels. These images provide a unique opportunity to identify biological mechanisms at a single-cell resolution but, at the same time, pose an important data analysis challenge, given the amount of data to be processed. This process can become especially overcomplicated when identifying common patterns shared across different patients' WSI images. A pipeline to help identify the most important regions to analyze in these images could improve the speed and accuracy of analysis.

Previous studies have proposed the usage of machine learning models in WSI images to predict regions-of-interest (ROIs)[21]. However, most of these models are developed for hematoxylin and eosin stain (H&E) images and trained for the task of ROI identification based on manual annotations of pathologists. This creates obstacles for the analysis of t-CycIF image data: A) Models trained for H&E images typically expect three channels as input and can not handle images like those produced by t-CycIF B) the manual annotations of the ROIs are not the most common label to have for these type of images and therefore might not be available; and C) the model is trained to identify the same regions that the pathologist had previously identified, therefore, not aimed for discovering unidentified patterns. A tool that uses t-CycIF as an input and can identify ROIs related to the biological processes of interest, such as therapy resistance, in a less guided fashion might help understand important mechanisms that were not previously identified.

ROIs can be very helpful in a clinical setting, by highlighting important regions to pathologists, speeding up their daily practice. In a research setting, ROIs related to process of interest, as it could be therapy resistance, might guide researchers to novel discoveries.

A joined data analysis pipeline that uses both t-CycIF and spatial transcriptomics technologies can have great potential for biological discoveries. Some spatial transcriptomics technologies, such as GeoMX can allow the identification of tumoral regions at cellular resolution. While t-CycIF can be utilized for whole slide images, GeoMX allows the maximum size of the region to be analyzed of 660 µm × 785 µm. In the aforementioned data analysis pipeline, an ROI from the whole slide image must be selected to use the sample for an experiment with GeoMX technology. The proposal of identifying ROIs with machine learning models can greatly help enhance the processing time with GeoMX and provide meaningful areas that may guide new discoveries. At the same time, GeoMX spatial transcriptomics results can be used to validate that the proposed ROIs are biologically meaningful.

Existing single-cell segmentation and cell phenotyping methods [27, 8] yield impressive results in characterizing individual cells, but important spatial information is lost in the single-cell-level analysis. Important tumor microenvironment (TME) structures such as tertiary lymphoid structures (TLS) [8] are better identified with a spatial context. Therefore, I propose an end-to-end pipeline that uses the t-CycIF images with minimum preprocessing, to train a model that produces ROIs related to classification tasks.

1.1 Rationale

To interpret whole-slide images and make biological conclusions, it is required to identify the most important ROIs for downstream analysis, such as when analyzing the tissue with spatial transcriptomics technology. Existing AI methods cannot process t-CycIF highly multiplexed imaging, as they are designed for 3-channel H&E images. Solving this can allow the identification of new biological mechanisms and improve patient stratification.

1.2 Aim

In this work, I aim to

1. develop a preprocessing method to input large-scale whole-slide multichannel immunofluorescent images to a machine learning classifier pipeline.

2. develop a machine learning pipeline to classify images based on clinical features, such as survival and therapy resistance
3. from the classifier in the previous step, extract ROIs used by the classifier that are relevant to biological processes

2. Background

2.1 High Grade Serous Ovarian Carcinoma

High Grade Serous Ovarian Carcinoma (HGSOC) is the most common type of ovarian cancer. It is responsible for 85% of the fatalities from all forms of ovarian cancer [42]. The 10-year survival rate for patients with early-stage HGSOC is 55%, while it is only 15% for patients with an advanced-stage disease [37]. One of the principal reasons for the elevated mortality of HGSOC is that it is most commonly diagnosed in advanced stages [28]. Only 13% of HGSOC cases are diagnosed at stage 1 or stage 2. A better understanding of the biology of HGSOC has led to the identification of new treatment options for a subset of patients. Genetic alterations cause distinct patterns in the development and progression of HGSC, with an effect on clinical behaviors. The samples can be categorized based on the status of a DNA-repair Homologous Recombination (HR). HR-deficient (HRD) samples accumulate high chromosomal abnormalities, including loss of heterozygosity, large scale transitions and telomeric allelic imbalances, which are indicative of high HRD score. In total, around 50% of the HGSC cases manifest HR deficiency, often resulting from alterations in tumour suppressor BRCA1/2 genes ([6],[40]). The identification of homologous recombination deficiency led to the successful targeting of these biological drivers and the discovery of synthetic lethality induced by PARP inhibitors (PARPi) proven to be effective at least in BRCA-mutated cancers [51]. Nonetheless, advanced HGSOC survival had a minor improvement over the last decades [53]. A better understanding of HGSOC mechanisms, and especially mechanisms of therapy resistance are needed to improve the overall survival of this deadly cancer.

2.2 Primary Debulking Surgery and Interval Debulking Surgery

Primary Debulking Surgery (PDS) is an approach where the patient first undergoes surgery to remove as much of the tumor as possible, followed by chemotherapy. PDS has been the standard of care for advanced ovarian cancer since the 1970s[10]. Interval Debulking Surgery (IDS) involves giving neoadjuvant chemotherapy (NACT) first to shrink the tumor, followed by surgery to remove any remaining disease, and then additional chemotherapy after surgery. IDS has emerged as an alternative to PDS in recent years[10]. The main difference is that with PDS, surgery precedes chemotherapy, while with IDS, chemotherapy is given first to shrink the tumor before surgery is performed. As the tissue analyzed with the t-CycIF pipeline is extracted during the surgery, PDS tissues are not exposed to chemotherapy, while IDS tissues are.

2.3 Tumor microenvironment

The tumor microenvironment (TME) is a complex and dynamic network that surrounds and interacts with cancer cells, playing a crucial role in cancer progression, metastasis, and treatment resistance. It consists of various cellular and non-cellular components, including immune cells, fibroblasts, endothelial cells, extracellular matrix (ECM) proteins, signaling molecules, and blood vessels. An illustration of the tumor microenvironment can be seen in Figure 2.1. Evidence across different cancer types shows the importance of TME in responding to a broad range of cancer therapies, further reinforcing the idea that the characterization of the tumor alone is not enough to understand therapy mechanisms and therapy resistance [44]. The tumor microenvironment is orchestrated by malignant epithelial cells and guided by tumor-induced interactions [52]. Tumor-derived signals downregulate anti-tumor functions from immune effector cells, resulting in tumor escape from the host immune system. Certain therapy strategies can be used to modulate the microenvironment towards one with anti-tumor activity [52]. In the case of HGSOC, distinct characteristics of the tumor microenvironment have been identified across different patients where characteristics of increased immunosurveillance were detected [25].

2.4 Biomarker-based therapeutics

Many existing approved FDA treatments require the identification of biomarkers for the selection of patients who can benefit from them. For example, in breast cancer treatment, anti-HER2 agents are dependent on identifying the patient's cancer as being HER2 positive [36], in the case of non-small-cell lung cancer (NSCLC), the treatment decision is highly influenced by several biomarkers [32] and in HGSC BRCA mutation and HRD status are important biomarkers for PARPi therapy [40]. More biomarkers are yet to be discovered. With existing technologies such as immunofluorescence imaging, single-cell analysis, and other omics data, there is an

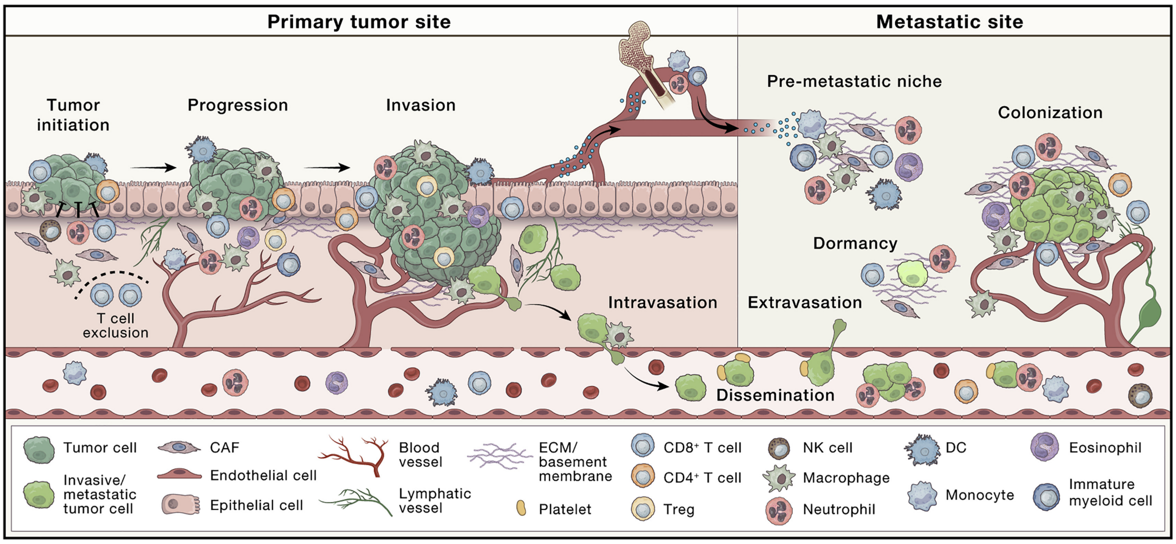


Figure 2.1: Tumor microenvironment illustration, showing the different types of cells and typical steps occurring in the microenvironment for a primary tumor site and a different metastatic site. Extracted from [48].

important challenge in finding patterns in amounts of data that are difficult to process at a human level. Machine learning could help in making sense of the vast amounts of omics data that we have and find new actionable biomarkers [11].

2.5 Tissue microarray

Tissue microarray (TMA) is a useful technique in the field of pathology, it helps to reduce the amount of data to be processed and reduce the costs of experiments, allowing the processing of multiple samples in parallel [16]. This technique can be useful for biomarkers discoveries using large patient cohorts. A microarray contains several tissue samples from different cases assembled on a single histologic slide. Each sample is extracted with a cylindrical shape from a bigger structure. With this technique, hundreds of samples can be arrayed into a single paraffin block. It allows simultaneous analysis of molecular targets under identical, standardized conditions on a single glass slide. It is an effective tool for high-throughput molecular analysis of tissues. An illustration of the process of building tissue microarrays can be seen in Figure 2.2.

2.6 Tissue Cyclic Immunofluorescence Imaging

Tissue-based cyclic immunofluorescence (t-CycIF) is a method for highly multiplexed immuno-fluorescence imaging. It can produce up to 60-plex images on an interactive

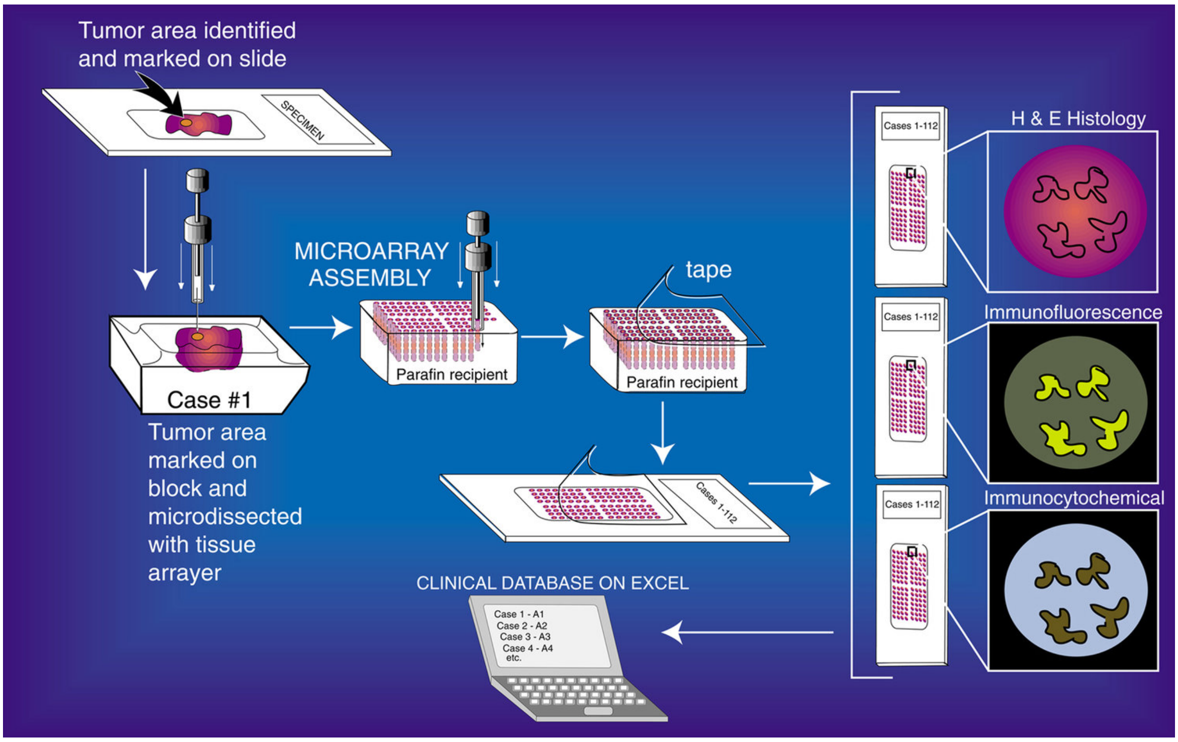


Figure 2.2: Diagram showing the steps involved in a tissue microarray extraction process and the resulting sampled material. Extracted from [2].

process, in which low-plex fluorescence images from the same sample are repeatedly collected and finally assembled into a high-dimensional image. High plex imaging can greatly expand the number of identified cell populations in the TME[29]. It is an effective method for pre-clinical and clinical research[29]. This method allows simultaneous visualization of multiple proteins within a tissue section and provides detailed maps of protein expression and localization, being a great tool to understand the TME further and for biomarker discovery.

2.7 Computer Vision

Computer vision is a field of artificial intelligence (AI) that enables computers and systems to derive meaningful information from digital images, videos, and other visual inputs and to take actions or make recommendations based on that information. Previously, it was considered that the goal of computer vision was to mimic human visual perception. Therefore, the robustness of a computer vision algorithm should be judged against the performance of a human performing the same task [34]. Deep learning methods are the state-of-the-art in several fields, including computer vision[49]. In the last decade, deep learning algorithms have been reported to exceed the human-level performance on some computer vision tasks [19]; this creates the possibility that machine learning algorithms could be applied to tasks in which humans have issues. Given the vast amount of data present in each t-CycIF image, computer vision could be a powerful tool to discover patterns that are difficult for the human eye to notice.

2.8 Machine learning for digital pathology

Whole slide images (WSI) of tissues can result in very large images, where a resolution of 100k x 100k is common [38]. Processing this amount of information by a human expert is not easy and, in general, requires that a pathologist select a region of interest to focus on. In a clinical setting, this creates subjectivity in the diagnosis tasks as different pathologists might select different ROIs and arrive at different conclusions [35, 41, 33] while in a research setting, there might be potential biomarkers that are showing patterns that are not easy to detect; machine learning can help develop more efficient pathology workflows, standardizing and automating manual and subjective tasks [38] and can help find new promising biomarkers. If that type of sample is run on a t-CycIF method, it can result in $100k \times 100k$ pixels by 60 channels, which is more than 600 billion pixels. This creates a huge challenge for human experts to identify relevant patterns. Machine learning can be critical in processing this data, finding patterns, and selecting ROIs that might lead to identifying novel biomarkers. At the same time, this amount of data can be challenging to process with a machine-learning pipeline. Therefore, different techniques have been developed to be able to process WSI images in a typical pipeline. For example, a common preprocessing step, is to divide a WSI into smaller tiles, so a group of tiles can be fitted into memory. This can be seen in Figure 2.3.

The state of adoption of machine learning in this field is heterogeneous in clinical practice. [13] provides a comprehensive analysis of the current status of different subareas in cancer pathology. A typical pipeline for deep learning in digital pathology is provided, which contains a series of steps that is common across most applications (can be seen in Figure 2.4). [13] classify deep learning models for digital pathology in two categories: basic, which aim to simplify routine workflows that are currently entirely performed by human pathologists (tumor detection, subtyping, and grading), and advanced (survival prediction, mutation, response prediction), which includes those that exceed pathologistsâ routine capacities. As part of the categorization provided by this article, it explains that basic systems could potentially reduce cost or turnaround time, but only in rare cases do they improve sensitivity and specificity as compared to human expert observers. On the other hand, it states that advanced models, with the inference of genetic alterations, prediction of survival and prediction of treatment

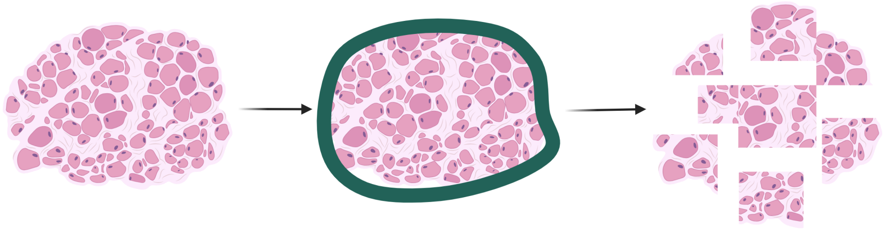


Figure 2.3: Typical tiling process, to divide WSI image into smaller images that could be fitted into a machine learning pipeline.

response, have the potential to change clinical decision-making in the management of solid tumours. The review article, which is from 2021, described that there is only one FDA-approved model for the basic case and none for the advanced one. In Figure 2.5, it is possible to visualize the landscape provided. For both categories, it is explained that a critical challenge is the validation that a model generalizes to different patient cohorts and is not overfitting to data from only one patient cohort used for model training. This suggests that the impact of this tool is still to be seen in daily clinical practice. As explained in the cited publication, H&E slides are routinely available for almost every cancer patient, making them an easy-to-obtain data source. Therefore, most of the existing work in digital pathology focuses on this type of images.

2.9 Biological interpretation of ML features

Deep neural networks are a type of machine learning model that is known for the difficulty to interpret and is commonly referred to as a black box [4]. There are multiple attempts to explain deep learning model decisions. In the case of computer vision, with algorithms to try to identify the pixels that caused the activation of a model [43] or model architectures that, by their design, provide interpretability information [12]. Nonetheless, proving the value of such interpretations from a biological point of view is not a solved task. In [26], they used a classifier model trained to predict response to platinum-based therapy in HGSC from H&E images and extracted the image regions used by the model for its prediction, which are then the selected regions to obtain the spatial transcriptomic (10X Genomics Visium) information from. After obtaining the spatial transcriptomics profile from the high-importance regions for the classifier model, the spatial transcriptomic was obtained from other regions (not important for the classifier model) and both transcriptomic profiles were compared. Tissue microarrays (TMAs) were constructed, and those were used for the spatial transcriptomic analysis. On those TMAs, differentially expressed genes were identified between short and long PFI groups. The inter and intragroup distances of the ST profile for the HC (high confidence) and BG (background) regions were measured. The intergroup distance is the cosine distance between long and short PFI, and the intragroup distance refers to the distance between background and high-confidence regions for the same group. To detect the most relevant alterations between short and long-survival groups, they identified the differential gene expression of the high-confidence regions of those groups. A gene set enrichment analysis was carried out. They found several enriched pathways in each of the long- and short-PFI groups. Important robust differences were identified in high confidence areas of both groups but not in background regions.

This example illustrates a path to get a biological validation of regions of interest proposed by a machine learning model. The model was trained to predict platinum-free interval (PFI) after at least 6 cycles of platinum-based chemotherapy, and it considered either 6 months or less (short PFI; PFI-S) or 18 months or more (long PFI; PFI-L) classes. An illustration of the pipeline used in the referenced article can be seen in Figure 2.6. The study used samples from Helsinki Biobank. All samples belong to HGSC patients with stage III or IV disease who underwent primary debulking surgery. As a limitation, they kept only two TMA cores for each sample (one high confidence and one background region), potentially losing valuable information. Another big limitation of the article is the selection of groups; as they only focused on very extreme cases of short and long survival, they ended up excluding most patients (from 517 patients, they only ended up including 52 patients). Furthermore, from those 52 patients, the sample size ends being reduced to 16 patients.

2.10 Variational Auto Encoders

Variational autoencoders (VAEs) are a type of generative neural network model that learns a probabilistic latent representation of input data [24]. VAEs encode inputs as a distribution in the latent space. VAEs can generate new samples by sampling from the learned latent distributions, enabling them to be used as generative models. VAEs consist of an encoder network that maps inputs to parameters of a latent distribution, and a decoder network that maps samples from the latent space back to the input space. The model is trained to maximize the evidence lower bound (ELBO), which encourages the latent distributions to capture the statistical structure of the training data. Autoencoders architectures have been used in the literature to learn a latent representation of digital pathology WSI images[14]. A typical VAE architecture can be seen in Figure 2.7.

2.11 ResNet classifier

Residual Networks (ResNets) are a class of deep convolutional neural networks introduced by He et al. in the 2015 paper, "Deep Residual Learning for Image Recognition"[18]. ResNet addressed the challenge of training very deep networks, which was historically hindered by the vanishing gradient problem[20] and the degradation of accuracy as the network depth increased. By introducing residual learning, ResNet allowed for the effective training of networks with significantly greater depth. The core innovation of ResNet is the use of "residual blocks," which include identity shortcut connections that bypass one or more layers. These shortcuts enable the network to learn residual mappings instead of directly learning the desired transformation. As a well-established and widely recognized architecture, ResNet provides a consistent and reproducible baseline for evaluating novel methods.

2.12 Foundation models' (for pathology)

Foundation models are a class of machine learning systems that are pre-trained on large-scale, diverse datasets and designed to be adapted for a wide range of downstream tasks. In the context of pathology, foundation models are pre-trained on extensive collections of histopathological images, often sourced from various tissue types, disease states, and staining techniques. In this work, I explored two foundation model's: UNI and GigaPath. UNI is a Vision Transformer computer vision model for digital pathology[9]. It was trained using over 100 million tiles from over 100 thousand H&E WSI slides and can produce embeddings over a given digital pathology image. Prov-GigaPath is a model pre-trained on 1.3 billion pathology image tiles in 171,189 whole slides. Prov-Gigapath has two main components, a tile encoder and a slide encoder that uses the embeddings generated by the tile encoder, with their coordinates from the original image, to produce a slide embedding. The main idea behind these models is that the pretraining that they had on vast amounts of data will help when these models are applied to a new downstream task.

2.13 Attention

The attention mechanism has been proposed and applied successfully in the context of Natural Language Processing ([30], [47], [15]). It is designed to selectively focus on the most relevant parts of a sequence when forming a representation of the input. Rather than treating all parts of the input equally, attention assigns weights to each part based on its relevance to the task. In most applications of attention-related architectures, the sequence is a sentence where each of the words obtains a different "attention score", giving the importance of the elements for the task at hand. Nonetheless, several computer vision applications have benefited from attention mechanisms. In the context of computer vision, the sequence is generated by dividing an image into tiles and creating a sequence of those tiles as how they are represented in the image. This has resulted in a performance boost for several computer vision tasks with large datasets, and state-of-the-art models for numerous famous computer vision tasks use an attention layer as part of their architecture [17].

2.14 Multiple-instance learning

Multiple-instance learning (MIL) is a type of supervised learning paradigm where the label is assigned to a set or bag of instances, opposite to having a label per instance as in a typically supervised paradigm[5]. The MIL assumption for a binary problem is that every positive bag contains at least one positive instance. The individual instance labels are not known during training. This paradigm adapts very well to a digital pathology setting, where, in most cases, a label exists for a complete image. Still, there are no labels assigned to separate sections of the image, where most of the image might not contain any signal of the label, but only a small part of it is related to the label (i.e., cancer subtype label)[14]. MIL architectures that use Attention layers provide as part of their architecture a method for extracting model feature importance over images, which has a direct application in digital pathology.

2.15 F1-score

The F1-score is a widely used performance metric in classification problems, particularly when dealing with imbalanced datasets. It is the harmonic mean of precision and recall, providing a single value that balances the trade-off between these two metrics. Precision measures the proportion of correctly predicted positive observations out of all predicted positives, while recall measures the proportion of correctly predicted positive observations out of all actual positives. By combining these, the F1-score captures the performance of the model in scenarios where both false positives and false negatives are critical. I chose the F1-score as the evaluation metric to compare the models in this study because it offers a robust measure for assessing classification performance in the presence of imbalanced class distributions. The formula for the F1-score is given as:

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

where precision is calculated as

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

and recall is calculated as

$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

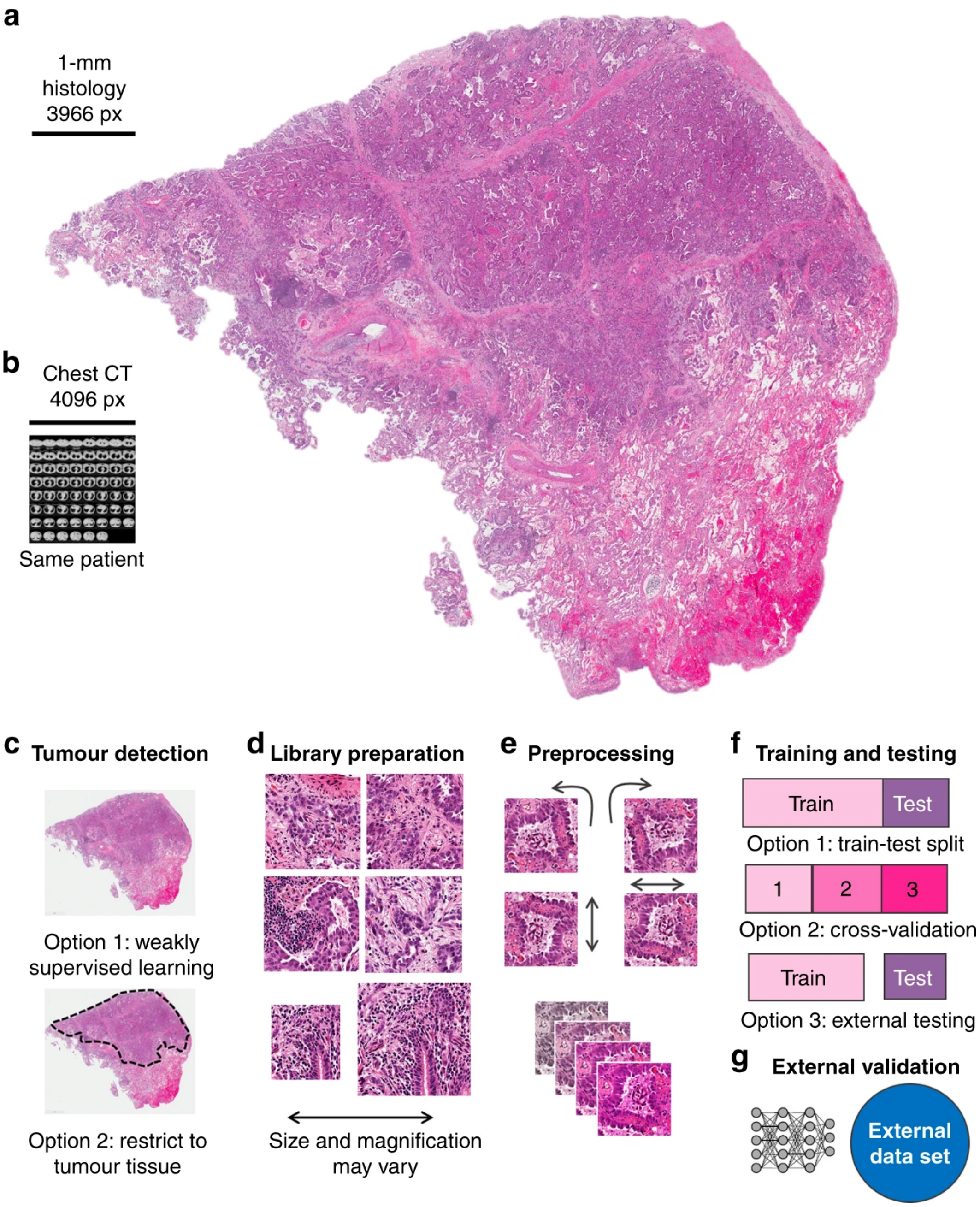


Figure 2.4: An example pathology deep learning pipeline, including preprocessing and training steps. Extracted from [13]

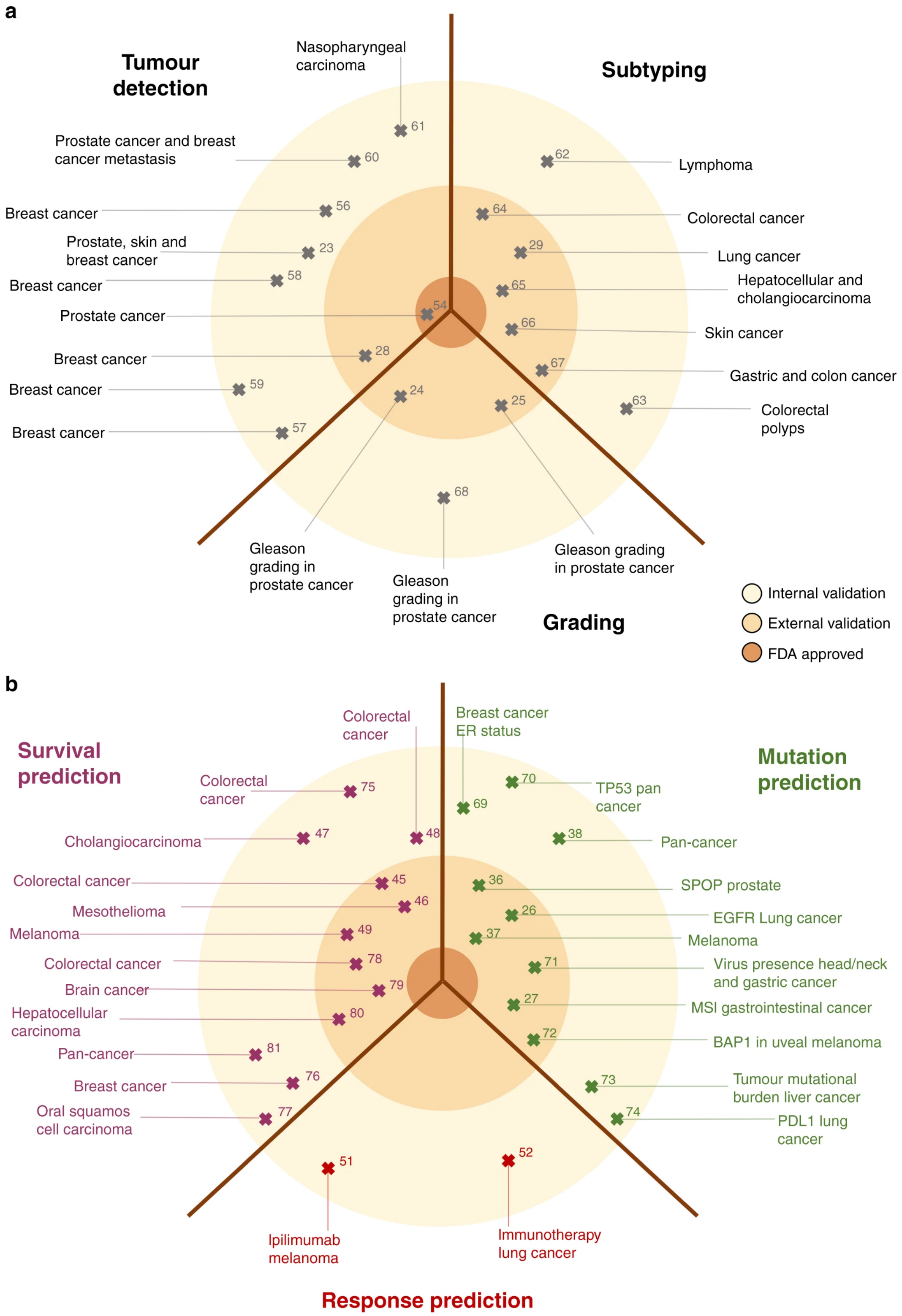


Figure 2.5: Existing models and validation status, divided in pipelines aimed to simplify basic pathology routine functionality (a) and those aimed to solve advanced pathology tasks capable of changing clinical decision-making (b), extracted from [13]

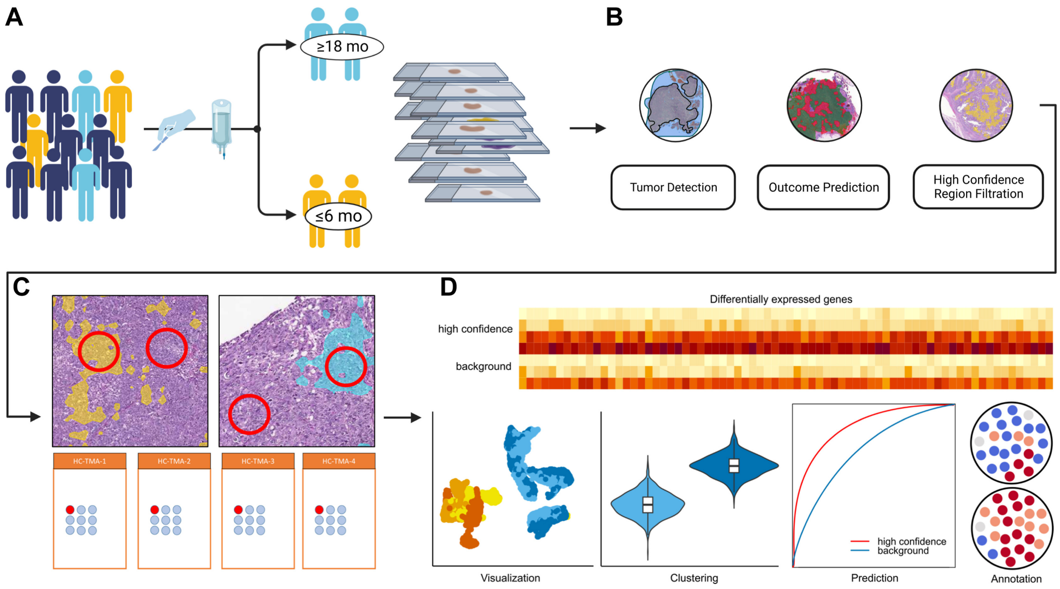


Figure 2.6: Description of pipeline used in [26] for exploring biological relevance of machine learning features, extracted from original publication.

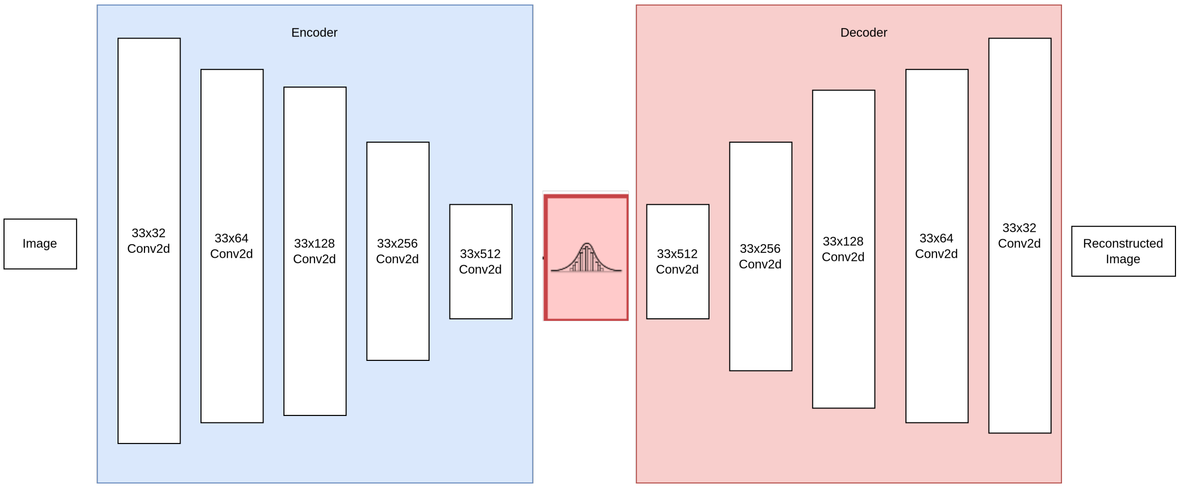


Figure 2.7: VAE architecture

3. Methods

3.1 Computational Analysis Pipeline

The method were implemented using Python programming language. Most of the analysis and model training was performed on computing clusters with access to graphics processing units (GPUs). As part of Aim 1), the images needed to be pre-processed, creating tiles out of these images with dimensions that can be easily used in the machine-learning pipeline. As part of creating the tiles, it was necessary to create a logic that detected the tissue position in the image and discarded the background. For Aim 2), the model implementation and training were carried out using the PyTorch Python library[39]. I tested three different classifier architectures: VAE combined with a simple convolutional network, ResNet, and Multiple Instance Learning, using a simple trained-from-scratch encoder and also pre-trained encoders for histopathology [54] based on the Vision Transformer [12] architecture, a computer vision implementation of the popular transformer architecture commonly used for natural language processing (NLP) tasks [46]. In the Multiple Instance Learning architecture that I used, which includes and Attention layer, an attention weights matrix can be easily computed to understand the regions of the image that the model assigns more importance to. I determined the suggested ROIs related to the classification task with this weights matrix. I adapted these existing models to multiplex imaging. Novel strategies were needed to handle the amount of channels available in this data.

3.2 Data

This work is based on highly multiplexed immunofluorescence images with at least 15 channels, which are contained in OME-TIFF [7] pyramidal files. These images contain single-cell-level information on the antibodies' expression in the immunofluorescence process. The proposed model will be trained on datasets from HGSC patients with FIGO stage IIb-IV. The number of channels and type of antibodies vary between the different datasets. Dividing the input images into tiles is necessary to accommodate the input data in dimensions the machine learning pipeline can handle.

3.2.1 NKI dataset

For tissue microarray (TMAs)[16] images, I used 29-channel (corresponding to different antibodies) 964 core images, with four cores per sample from 241 patients in Dutch tertiary referral hospitals (Netherlands Cancer Institute-Antoni van Leeuwenhoek Hospital (NKI-AVL), Maastricht University Medical Centre (MUMC) and Amsterdam University Medical Centre (AUMC)), between January 2008 and December 2015. A plot summarizing the dataset cohort can be seen in Figure 3.1. This data was collected initially by [45], which, in each case, sampled two cores from the tumor center and two more from the periphery. The TMAs were scanned with a RareCyte CyteFinder scanner following the t-CycIF protocol[29]. The image tiles were stitched and successively piled up with the corresponding images from every cycle to one another (registered) using the ASHLAR algorithm[3]. The TMA's cores were cropped from the stitched multi-channel image with the Image Processing Toolbox on MATLAB[1]; this process produced multi-channel tiff images.

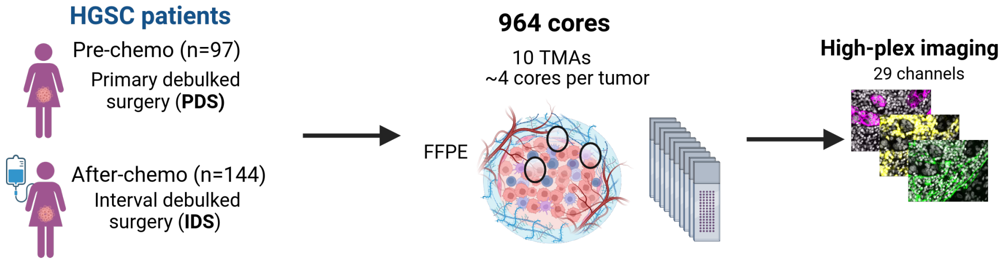


Figure 3.1: NKI dataset description. Created in https://BioRender.com.

3.3 Preprocessing: Normalization of t-CycIF data

The t-CycIF process implies using several cycles of low-multiplex imaging over the sample. As described in [29], highly multiplexed antibody intensities on the images are affected by the cycle number that they are run on. An example of this, with different antibodies, can be seen in Figure 3.2. For some antibodies, fluorescence intensity increases with cycle number, and for others, it decreases. The reasons for the variation in staining with cycle number are not known, but this variation is reproducible across specimens, which makes us believe that it reflects the properties of the antibody and not the t-CycIF process[29].

In this work, I created a normalization algorithm that normalized the intensity values based on the range they had for all images of the same dataset. With it, I changed the intensity values to be between 0 and 1, being one in the case that the value was at the highest of the values range for that dataset. Still, to the date of writing this work and my understanding, there is no normalization strategy that fully addresses this issue and that is a standard for this type of image.

Variation in staining can be controlled by staining all specimens with the same antibodies in the same order. Nonetheless, this solution will not work for the creation of large-scale t-CycIF datasets in which diverse tissues will be compared with each other and for which different datasets for different research questions get combined together. The intensity of this cycle-signal reduction varies between antibodies, with some more robust than others. It is a common practice to design the experiments with the most important antibodies related to the research question and hypothesis on the first cycles, as it is common that the tissue might be degrading in each cycle. Also, the datasets used for this work have an important imbalance between the classes used for the classifier model, which were initially collected for heterogeneous research questions. This can easily highlight data bias. For example, consider a dataset where 90% of the samples belong to class A and 10% to class B, and antibody X (which produces a weaker signal in later cycles) was measured in cycle 0. In contrast, another dataset with 90% of class B and 10% of class A measured antibody X in cycle 5. This setup might misleadingly suggest that antibody X has a stronger signal in class A. However, the apparent difference could actually be a bias introduced by the experimental design, rather than a true difference in antibody X signal between classes A and B. Therefore, it is important to have a normalization strategy for t-CycIF datasets. This could be an important obstacle if in the future a foundation model is trained on a diverse set of images from t-CycIF.

Examples of antibody signal variation in the dataset used for this work could be seen in Figure D.1 for NKI dataset dataset. In those plots, we see the frequency of pixel signals between 0 and 65535, corresponding to the TIFF images.

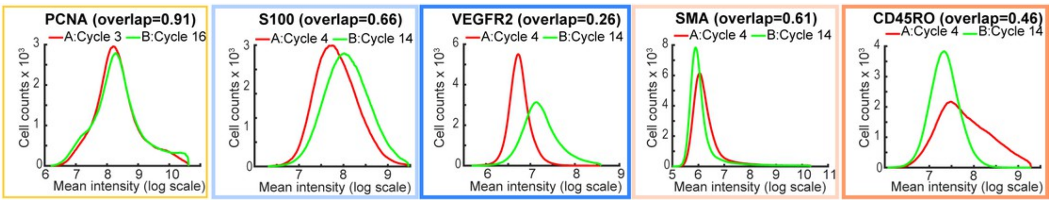


Figure 3.2: t-CycIF antibodies signal per cycle. Extracted from [29].

3.4 Cohort selection

On the TMA core images used for this work, as explained in section 3.2, there are four core images per sample. Two of those four images belong to the tumor center and two to the tumor periphery. These images were used to train a classifier to detect if a tissue was exposed to chemotherapy. It was determined that using the TMA core images from the tumor periphery might not be representative enough of the differences between samples exposed to chemotherapy, so it was tested to train the model, excluding the cores that did not contain enough cancer cells. For this purpose, a table containing cancer cells per core image was used, and cores containing less than 500 cancer cells were discarded. The table was created by first segmenting the cell nucleus, then the image Processing Toolbox on MATLAB was used to quantify the median intensity of each channel for the segmented cells, along with morphological features such as Area, Eccentricity, Solidity, Perimeter, and Roundness. After, the cells were clustered according to the z-score scaled signal of the epithelial markers (Cytokeratin 7 and E-cadherin) using a Gaussian mixture model implemented in the R-package Mclust. The number of Gaussians was fit for two, and the cells within the cluster of higher expression of epithelial markers were labeled as cancer cells.

3.5 Preprocessing: Tiling

Some of the WSI image files used in this project are larger than 100 GB. This amount of data exceeds what can be fitted in memory in the most commonly available graphics processing unit (GPU) used for machine learning training. This requires reducing the data size before inputting it into the model. An option for reducing the image size might be to rescale the image into a lower resolution, but doing so will imply that valuable information will be lost. Another alternative I follow in the present work consists of dividing the WSI images into multiple smaller files (tiles), which can then be fitted into the GPU memory.

An optimal process of tile creation requires 1) the identification of the tissue in the image (to avoid creating tiles that only contain background), 2) establishing a maximum acceptable background threshold to generate tiles on the borders of the tissue, and 3) defining a tile size. This process can be visualized in Figure 3.3.

To carry out this task, I adapted some of the methods proposed by [23], originally designed to run with H&E images, to the t-CycIF data. This package expected input images with three channels (from H&E images). In this case, data exceeds three channels, so deciding which channels should be used for the tiling process was needed. I chose to use staining of cell nuclei, 4’,6-diamidino-2-phenylindole (DAPI) for this as

DAPI allows the identification of the cell presence in t-CycIF data. Cell presence is necessary for detecting any biological signal related to any of the other channels in our data and, therefore, constituting an optimal guide for the tiling process. While in the original Histoprep implementation, tiles are saved as 3-channel images, in the algorithm that I created for this work, after deciding on the tiling coordinates, each tile was saved as a TIFF image, keeping all the channels that originally existed in the t-CycIF WSI image.

To maximize the signal obtained from each image, I follow the tiling procedure both with WSI images and with TMAs[16] images. In present work, I used TMA images with an approximate resolution of 3500x3500 pixels; although these images are much smaller, resizing them to a reasonable input size for a neural network model will still imply data loss. Therefore, I follow the same tiling procedure as for WSI images, which consists of first detecting the tissue location (ignoring the background) and splitting the image into smaller tiles. The core images I have available were not perfectly cut for each core, but most of them have a core in the center, with other cores invading the image's borders, as can be seen in Figure 3.4. This could cause several issues, as it is intended that the model learns a common pattern related to the class to predict what is present in the image. The labels are assigned to the core in the center, but the other cores invading the image might contain different labels. This could lead the model to base its predictions on a pattern existing in the invading cores and not on the main label assigned to the image. To mitigate the risk of including invading tissues in the image, I decided to crop the center of the image, which proved to be reliable enough in most cases to avoid the invading cores. A systematic search was conducted to determine the optimal cropping resolution, balancing the need to avoid including non-relevant areas while preserving the signal from the main core. Based on this search, a resolution of 2300x2300 pixels was selected for center cropping the images.

3.6 Classification task

The primary classification task for the trained models was to determine whether the tissue was extracted during PDS or IDS (see section 2.2). As IDS tissues were exposed to chemotherapy, the idea is that the effect of chemotherapy on the tissue should cause a clear pattern, different than PDS, which the model could learn. For the first simple comparison between models, data was manually split between train and validation. Validation split consisted of two whole slides from subsection 3.2.1 dataset, `TMA_45_312` and `TMA_46_325`. These two slides were selected for the validation set as they have a balanced number of TMA cores that were PDS and a similar number of IDS. The

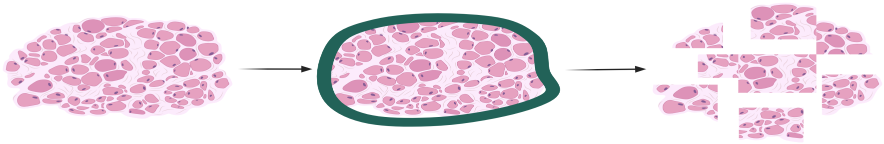


Figure 3.3: Tiling process: Detection of tissue area, and division of tissue image into smaller tiles

rest of the cores (from other slides) were used for the training set. Afterward, as a more robust comparison, all models were run on a random 5-fold cross-validation, were the whole slide, containing several TMA cores, was randomly put in either a train or validation group, to avoid data leakage and ensure that two cores from the same slide could not belong to different groups.

3.7 Model Comparisson

In this project, several model types were tested: variational autoencoder, ResNet, and MIL.

I trained VAE networks to try to capture a meaningful latent representation. The latent representations were later used as the input of a classifier for PDS vs IDS.

ResNet architecture was used as a baseline. This classifier gets as input a raw t-CycIF image from a TME core with three channels and more (several combinations were tested), which is resized to 512 by 512 pixels and was trained from scratch (without any pertaining) to classify PDS vs IDS. This serves as a base benchmark to compare any other solution.

For applying MIL to this work, using the tiled version of the core images, which in this case resulted in the bag of instance, I generated embeddings using pre-trained pathology models (section 2.12) and an encoder that was trained from scratch. Then, I used those embeddings as the input of a MIL classifier model with an attention layer based on the architecture proposed by [22]. An illustration of the pipeline steps can be seen in Figure 3.5. The use of an attention layer in this architecture provided a built-in method to explore the importance of the model assigned to each tile, which could be used to select ROIs. As part of the, pretained encoders, I tested the UNI model [9] and Prov-GigaPath [54] (tile encoder, Prov-GigaPath also has a slide encoder,

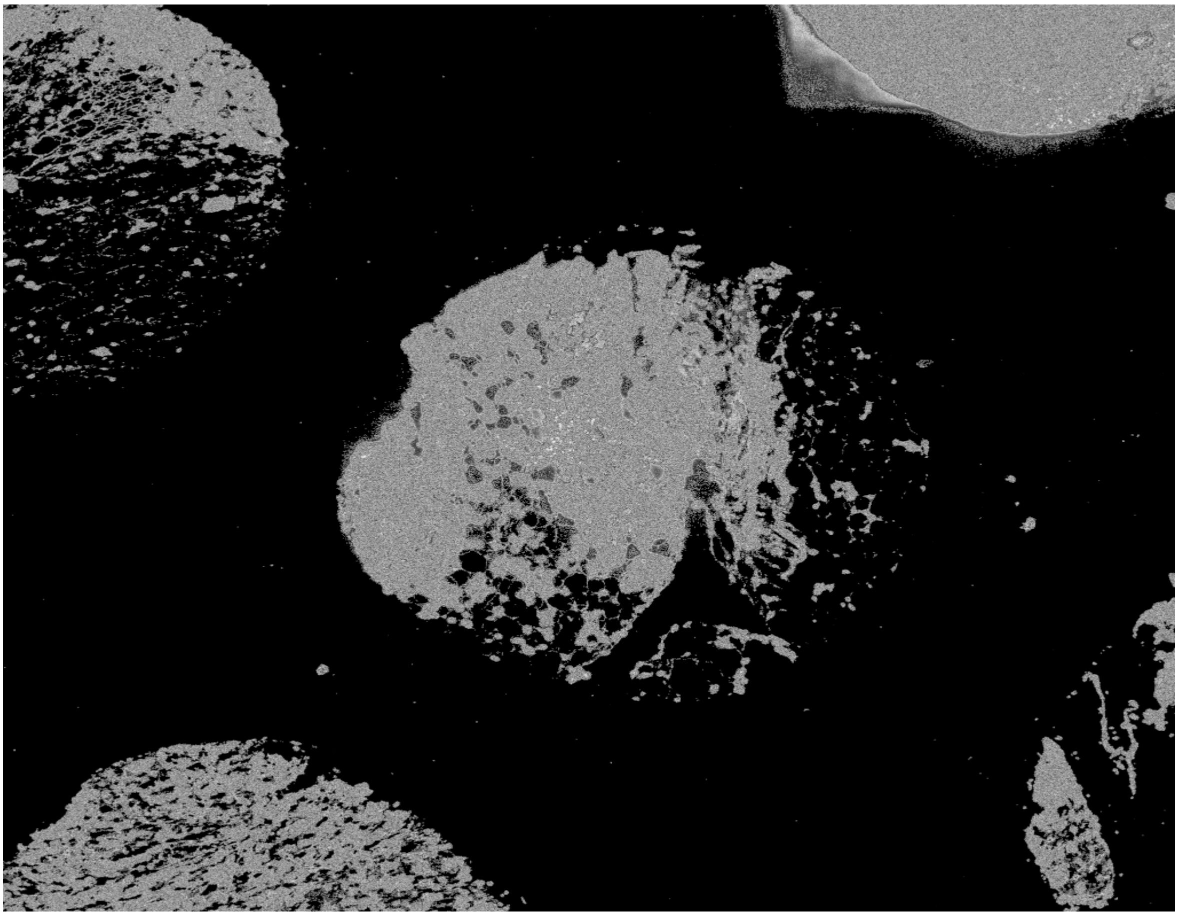


Figure 3.4: TMA image, with a center core with other cores invading the image

which was not utilized). Although both of these models were trained for H&E images, I adapted them to the highly multiplex data obtained from t-CycIF. These models expects three channels as input. To utilize the t-CycIF data, I made several three-channel combinations of all the data available in these images. After, the resulting embeddings were fed to different models for classification tasks.

3.8 Attention weights

Visualizing attention layer weights in a vision transformer model can provide insights about the model decisions and which parts of the input image it focuses on for a given task[12]. While some explainability methods, such as Grad-CAM[43] propose an algorithm to obtain explanations for model predictions, the Attention[47] layer is part of the model architecture and at the same can be inspected to obtain the importance values assigned to different sections of the input. This was used with the MIL model (section 2.14) to obtain ROIs.

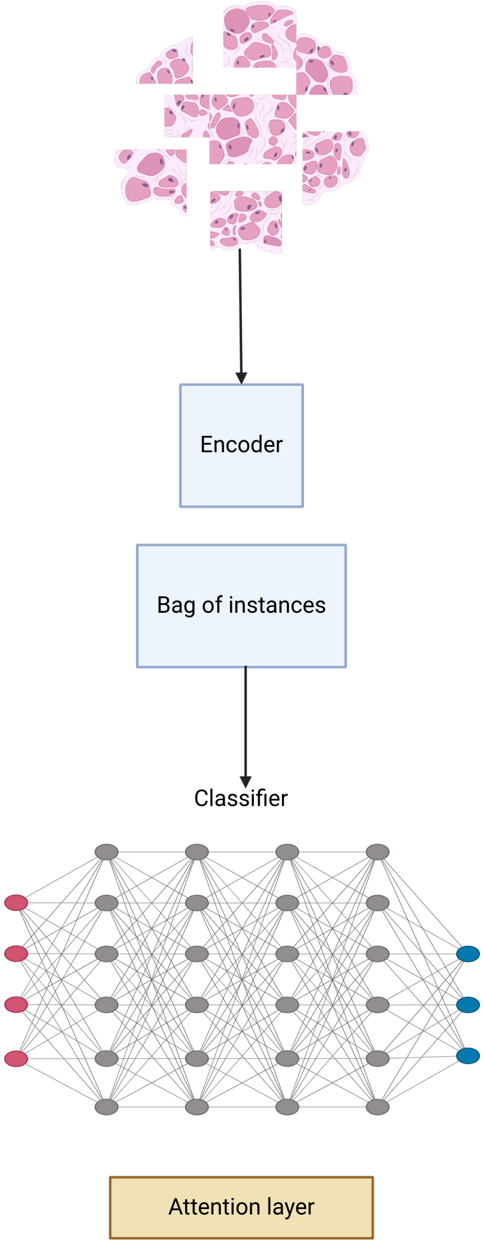
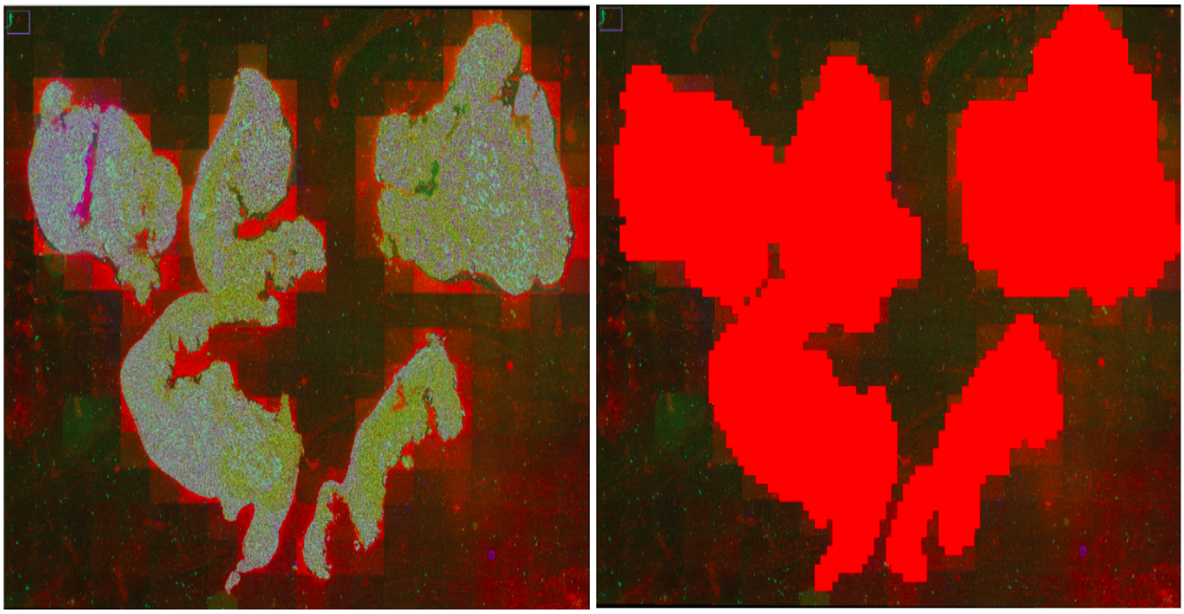


Figure 3.5: MIL pipeline: Encode each tile into a latent representation, construct bag of encodings with a common label as the input of a classifier with an attention layer

4. Results

4.1 Tiling

Results of the tiling algorithm can be visualized in Figure 4.1 for a WSI image and in Figure 4.2 in a TMA core.

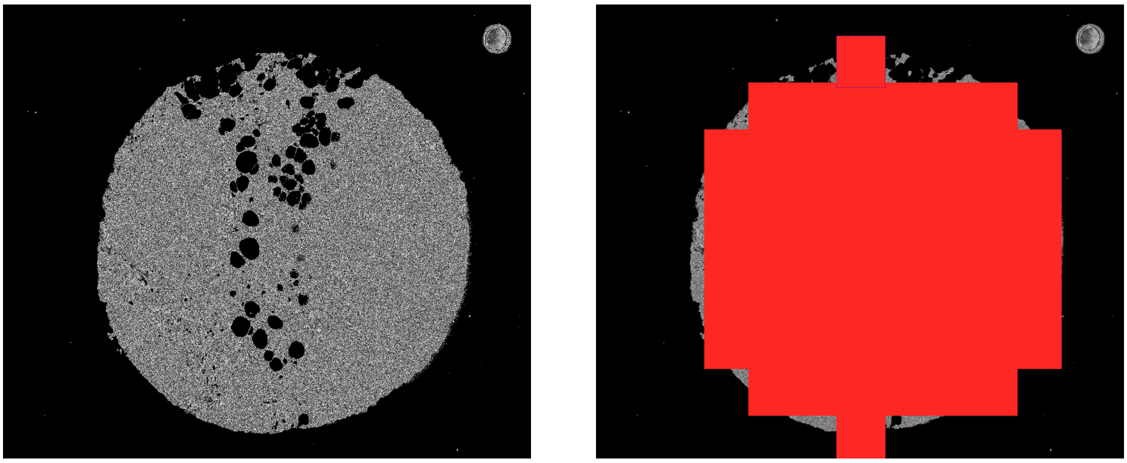


(a) Original image **(b)** Proposed tiling by algorithm

Figure 4.1: Detection of tiles over WSI image

4.2 Variational Auto Encoders

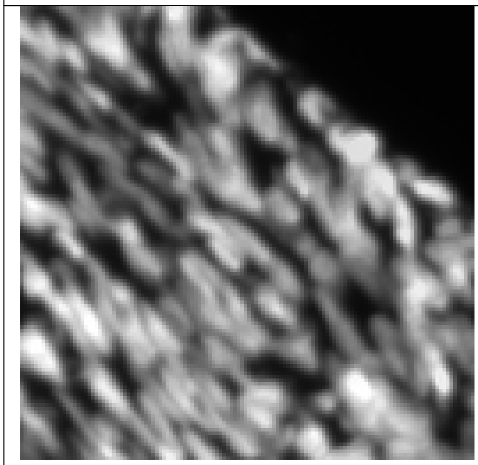
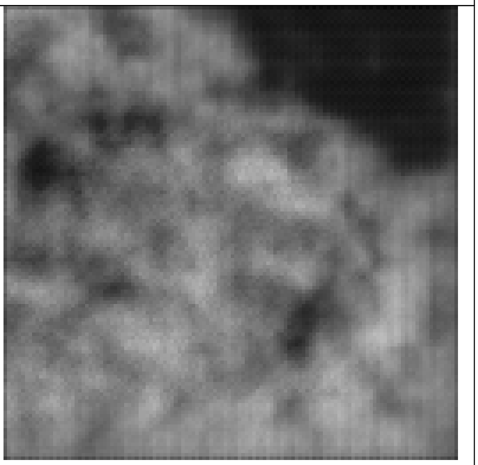
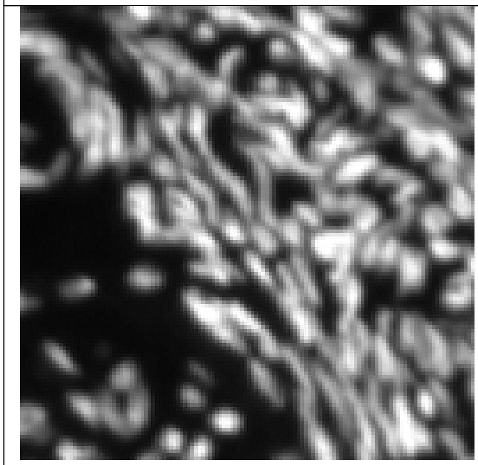
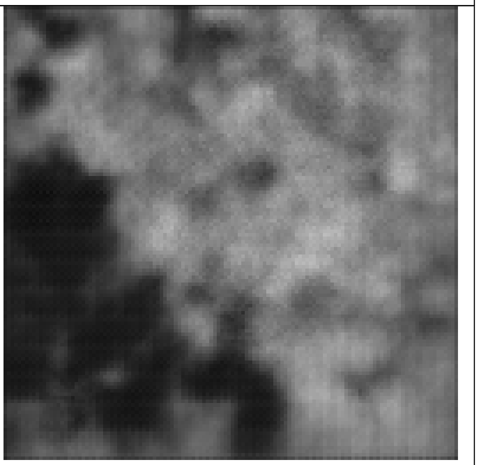
On the VAE results, good reconstruction quality was obtained over the original images, as it can be visualized in Table 4.1. However, I tested using the learned latent space as the input of a two-layer CNN classifier to predict if an image came from PDS or IDS and using TMA cores. The resulting performance was worse than using a ResNet with the full image as an input, as seen in Table 4.2. When visualizing the learned latent space with TSNE on 2D, no clear group separations are showing in the data, which



(a) Original image **(b)** Proposed tiling by algorithm

Figure 4.2: Detection of tiles over TMAs image

seems to imply that the learned latent representation might not be useful [31]. The learned latent space visualization can be seen in Figure 4.3.

Original image	Reconstructed image
	
	

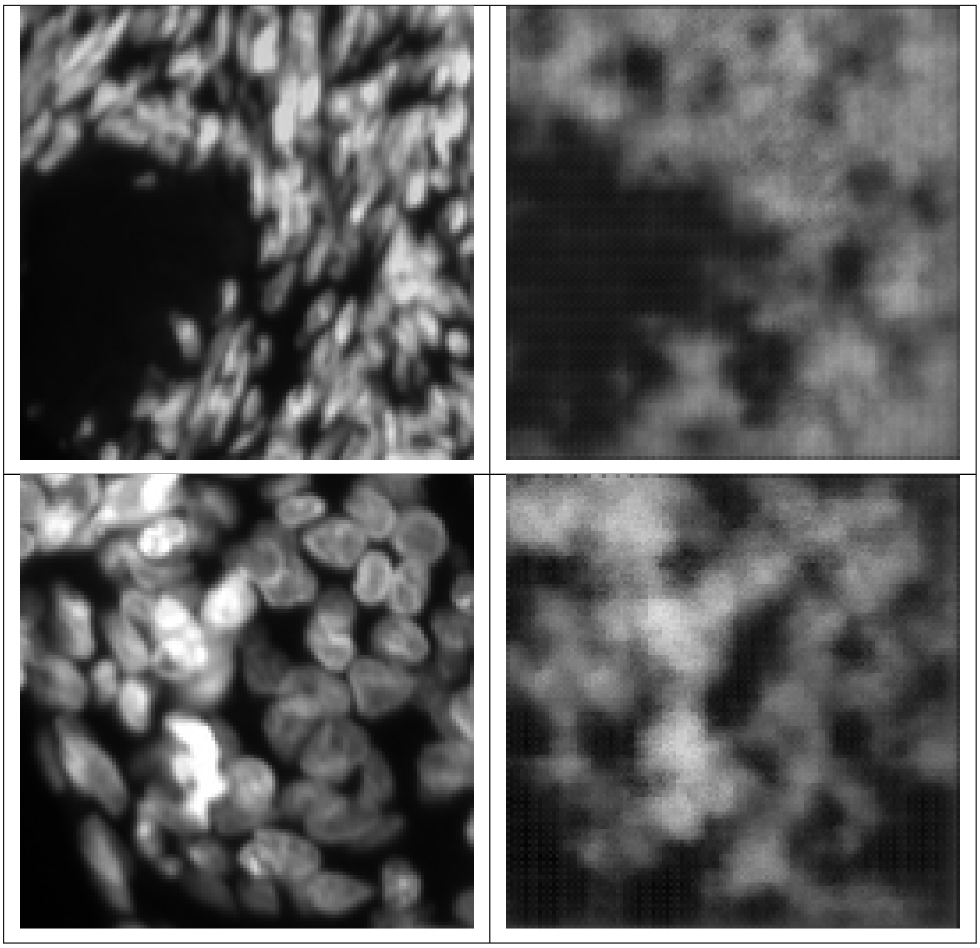


Table 4.1: VAE results, with original images and the reconstructed images

Architecture	Training set F1-score	Validation set F1-score
VAE latent space + CNN	0.66	0.49
ResNet	0.79	0.63

Table 4.2: PDS and IDS classifier performance

4.3 ResNet classifier

The trained ResNet classifier was tested on TMA core images, as this model was adapted to use raw images resized to 1024x1024px. This proved to keep enough signal when resizing TMA cores (that were originally between 2 and 4 times the size of the

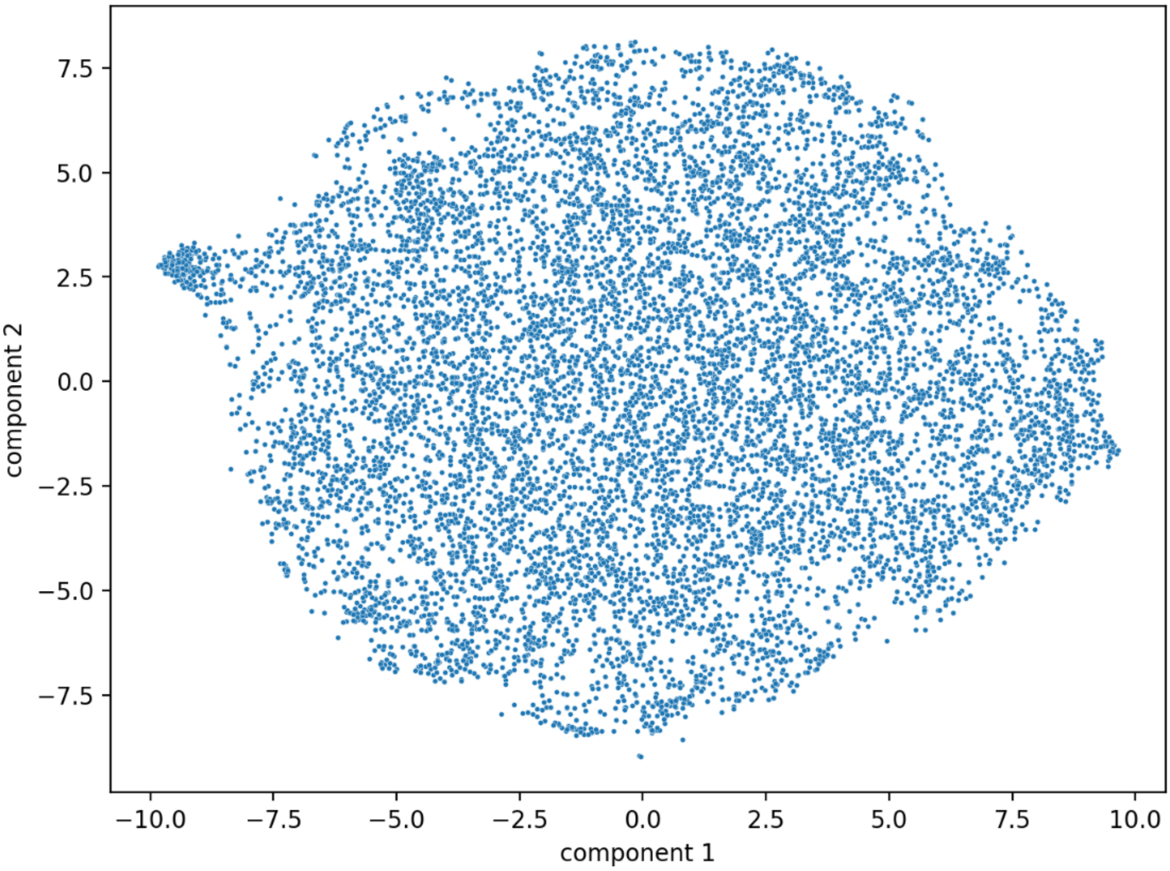


Figure 4.3: VAE latent space TSNE

resizing). The performance of this model was relatively good for predicting PDS and IDS therapy, achieving an F1-score of 0.63 on the validation set. An explainability algorithm must be used on top of the trained model to obtain ROI regions from this type of classifier. In this case, I used Grad-CAM[43]. An example activation can be seen in Figure 4.4.

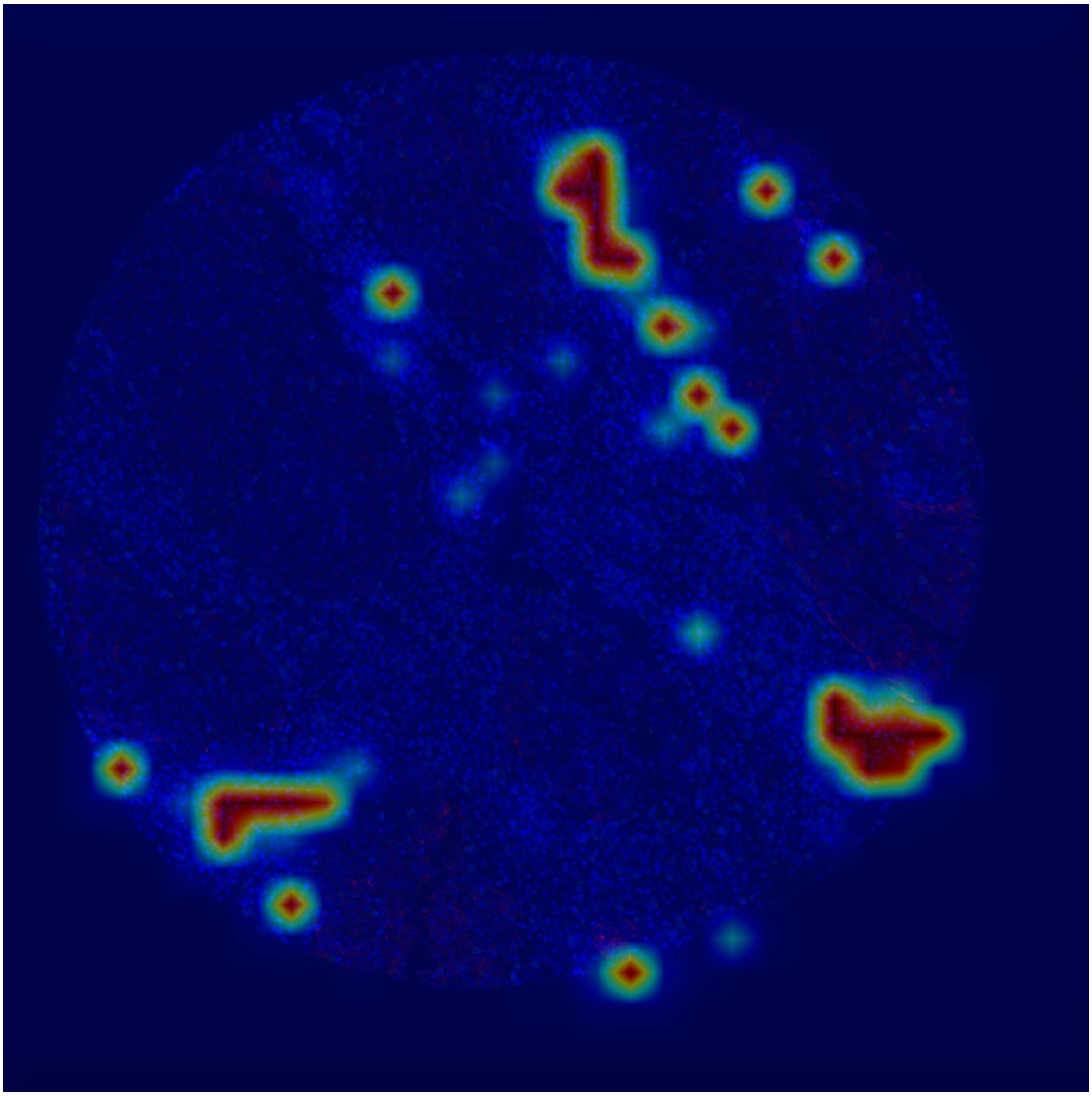


Figure 4.4: ResNet classifier: Grad-CAM activations

4.4 MIL classifier

The MIL classifier got the best results of the models tested on TMAs cores images. For the MIL pipeline, which is described in Figure 3.5, I tested training a simple convolutional neural network to encode the tiles information and also I tested using

UNI[9] and Prov-GigaPath[54] as the encoder model. In this pipeline the input are the image tiles, which are generated with the algorithm explained in section 3.5. Then, these tiles, selecting for certain channels (in this case DNA1, CK7, and Vimentin), are fed to either a trainable encoder part of the model or to a pre-trained model (UNI or GigaPath), and the resulting encoding is used as the input of an Attention-MIL classifier. The model returns the output of a Sigmoid layer for a binary classification. With the different encoders, as it can be seen in Table B.1, results from each of these variations it is quite similar. Both of these pre-trained encoders are designed to encode H&E images, which is very different from the t-CycIF data being used in this work. My hypothesis is that given the pre-training data differences, these foundation models are not especially useful out of the box and can be easily replaced with a simple encoder that have been exposed to a much smaller amount of data. As described in section 3.3, t-CycIF data has unique characteristics in terms of signal differences per antibody and cycle. This feature is not comparable to H&E data characteristics, and it is not clear how it can affect the encoders that were pretrained with this data type. Nonetheless, there is a slight performance boost with those pre-trained encoders, and I ran the rest of the experiments with a UNI and GigaPath encoder when using a MIL model. In Table 4.3, we can see that the best-performing model is the combination of Prov-Gigapath with a MIL classifier and the channels DNA1, CK7, and Vimentin. I ran the best performing models with a different combination of channels, which returned worse results. This can be seen in Table B.5.

DNA1, CK7 and Vimentin, 5-fold cross validation	**AVG F1 Score on Validation Set**
VAE + Classifier	0.56
ResNet	0.73
UNI encoder MIL	0.88
Prov-Gigapath encoder MIL	0.89

Table 4.3: Model Performance on Validation Set, with cross-validation

4.5 Identified ROIs

My hypothesis is that the model with the best performance should return more meaningful ROIs, and in addition, the MIL classifier, as in the architecture used in the present work, contains an attention layer, provides a built-in logic to extract the ROIs, which we assume that should be less biased than extracting these regions with an additional method, as it could be Grad-CAM[43]. Therefore, I extracted the ROIs produced by the combination of Prov-Gigapath with a MIL classifier and the channels DNA1, CK7, and Vimentin. In Figure 4.5,Figure C.1, Figure C.2,Figure C.3 and Figure C.4 we can see the detected regions of interest for the different classes (stronger red color means more weight of the model on the highlighted region).

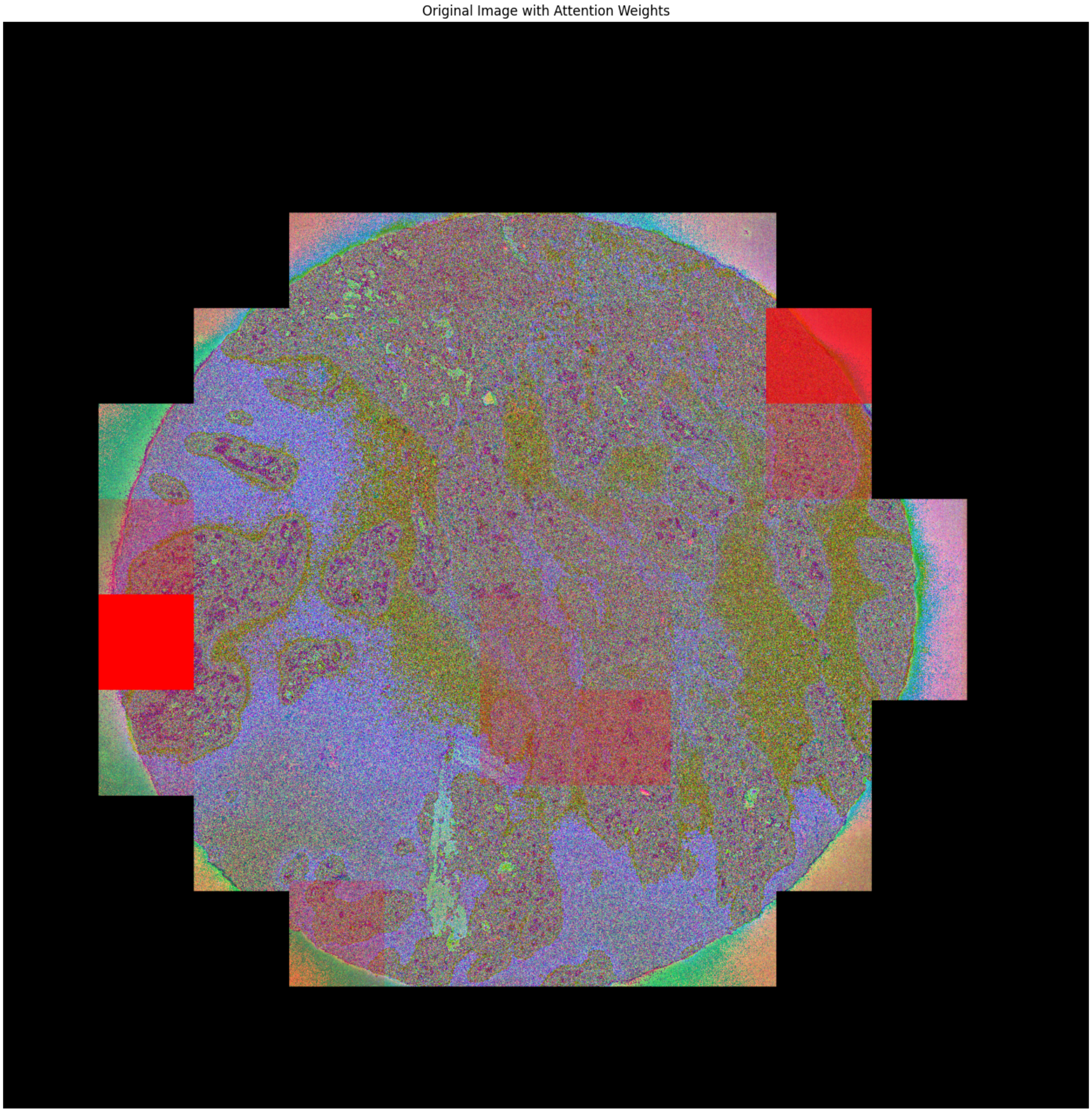


Figure 4.5: ROI sections in red, by analyzing attention layer, on a TMA exposed to chemotherapy

5. Discussion

In this work, I trained machine learning models to classify between tissues that had been exposed or not to chemotherapy, by classifying PDS and IDS patients section 2.2 using immunofluorescence images (t-CycIF) as input. The main opportunity in this data relies on the availability of extra information compared to the typical H&E pathology images. In the H&E images, there are three channels of information (RGB); in this case, we have between 25 and 50 channels. Nonetheless, in my results, I did not manage to train a model that used more than three channels in an effective way. In my opinion, this has multiple causes: A) Most publicly available pre-trained models for pathology data are pre-trained over H&E images with three channels. Using this kind of model with more than three channel inputs requires important architecture change, while the benefit of the pre-training might be lost due to the data input differences. B) Training a computer vision model from scratch, without any pre-training data, requires an important amount of images. The t-CycIF datasets available for this work are all of a limited size. Combining multiple datasets in the context of t-CycIF creates several challenges, as most datasets do not follow the same combination of channels and their order. This creates important variation in input data as explained in section 3.3. From the tested models, Multiple Instance Learning models section 2.14 combined with pre-trained pathology models, performed the best, showing promising results. On the other hand, neither a classifier based on the input of a VAE nor the ResNet classifier showed significant results. As far as I am aware, no other work has explored the use of pre-trained H&E models in immunofluorescence data. My results have shown that these pre-trained models, although for a different data modality, outperformed an encoder trained exclusively with immunofluorescence data and benefitted from the number of images that the pre-trained models were exposed to.

6. Conclusions

This work demonstrates the feasibility of applying machine learning models to t-CycIF immunofluorescence images for classifying ovarian cancer tissues and identifying biologically relevant regions of interest (ROIs). Among the models tested, the MIL classifier with pre-trained encoders (UNI and Prov-GigaPath) achieved the best performance, suggesting that attention-based mechanisms can effectively highlight critical tumoral areas and that the pre-training of these models in a different data modality (H&E) provides useful information for these models when applied to immunofluorescence images. The identified ROIs provide valuable insights into the tumor microenvironment and potential therapy resistance markers.

Despite these promising results, challenges remain in leveraging the full potential of t-CycIF multi-channel data. Existing foundation models are optimized for 3-channels H&E-stained images, limiting their direct applicability to highly multiplexed imaging. Additionally, data normalization across different t-CycIF datasets remains an open problem, impacting model generalizability.

6.0.1 Future work

New normalization strategies are required for effectively combining immunofluorescence datasets. With the normalization solved, a computer vision foundational model trained over a vast immunofluorescence dataset might be the key to fully take advantage of the rich information present in immunofluorescence images.

Use of AI tools

Grammarly (https://app.grammarly.com) has been used to check grammar. ChatGPT (https://chatgpt.com/) has been used for proofreading and grammar checking.

Acknowledgments

I want to thank my supervisor, Anniina Färkkilä, for her guidance and encouragement throughout my work.

I am grateful for the opportunity to work and gain valuable experience in Färkkilä's research lab, and for her to believe in me and guide me on my journey.

I want to thank Fernando Perez, Oskari Lehtonen, Ada Junquera, Angela Szabo and Ziqi Kang for their technical assistance, as well as for their constructive feedback and insightful discussions. I would like to thank everyone in Färkkilä Lab for making me feel at home in the office, and for all the discussions we had about science.

Lastly, I am grateful to the Jura community for its invaluable support.

Bibliography

[1] MATLAB.

[2] How to Construct Tissue Microarrays, Sept. 2017.

[3] labsyspharm/ashlar, July 2024. original-date: 2017-05-15T11:19:52Z.

[4] J. P. Amorim, P. H. Abreu, A. FernÃ¡ndez, M. Reyes, J. Santos, and M. H. Abreu. Interpreting Deep Machine Learning Models: An Easy Guide for Oncologists. *IEEE Reviews in Biomedical Engineering*, 16:192–207, 2023. Conference Name: IEEE Reviews in Biomedical Engineering.

[5] B. Babenko. Multiple Instance Learning: Algorithms and Applications.

[6] D. Bell, A. Berchuck, M. Birrer, J. Chien, D. W. Cramer, F. Dao, R. Dhir, P. DiSaia, H. Gabra, P. Glenn, A. K. Godwin, J. Gross, L. Hartmann, M. Huang, D. G. Huntsman, M. Iacocca, M. Imielinski, S. Kalloger, B. Y. Karlan, D. A. Levine, G. B. Mills, C. Morrison, D. Mutch, N. Olvera, S. Orsulic, K. Park, N. Petrelli, B. Rabeno, J. S. Rader, B. I. Sikic, K. Smith-McCune, A. K. Sood, D. Bowtell, R. Penny, J. R. Testa, K. Chang, C. J. Creighton, H. H. Dinh, J. A. Drummond, G. Fowler, P. Gunaratne, A. C. Hawes, C. L. Kovar, L. R. Lewis, M. B. Morgan, I. F. Newsham, J. Santibanez, J. G. Reid, L. R. Trevino, Y.-Q. Wu, M. Wang, D. M. Muzny, D. A. Wheeler, R. A. Gibbs, G. Getz, M. S. Lawrence, K. Cibulskis, A. Y. Sivachenko, C. Sougnez, D. Voet, J. Wilkinson, T. Bloom, K. Ardlie, T. Fennell, J. Baldwin, R. Nichol, S. Fisher, S. Gabriel, E. S. Lander, L. Ding, R. S. Fulton, D. C. Koboldt, M. D. McLellan, T. Wylie, J. Walker, M. OâLaughlin, D. J. Dooling, L. Fulton, R. Abbott, N. D. Dees, Q. Zhang, C. Kandoth, M. Wendl, W. Schierding, D. Shen, C. C. Harris, H. Schmidt, J. Kalicki, K. D. Delehaunty, C. C. Fronick, R. Demeter, L. Cook, J. W. Wallis, L. Lin, V. J. Magrini, J. S. Hodges, J. M. Eldred, S. M. Smith, C. S. Pohl, F. Vandin, E. Upfal, B. J. Raphael, G. M. Weinstock, E. R. Mardis, R. K. Wilson, M. Meyerson, W. Winckler, G. Getz, R. G. W. Verhaak, S. L. Carter, C. H. Mermel, G. Saksena, H. Nguyen, R. C. Onofrio, M. S. Lawrence, D. Hubbard,

S. Gupta, A. Crenshaw, A. H. Ramos, K. Ardlie, L. Chin, A. Protopopov, J. Zhang, T. M. Kim, I. Perna, Y. Xiao, H. Zhang, G. Ren, N. Sathiamoorthy, R. W. Park, E. Lee, P. J. Park, R. Kucherlapati, D. M. Absher, L. Waite, G. Sherlock, J. D. Brooks, J. Z. Li, J. Xu, R. M. Myers, P. W. Laird, L. Cope, J. G. Herman, H. Shen, D. J. Weisenberger, H. Noushmehr, F. Pan, T. Triche Jr, B. P. Berman, D. J. Van Den Berg, J. Buckley, S. B. Baylin, P. T. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, L. R. Jakkula, S. Durinck, J. Han, S. Dorton, H. Marr, Y. G. Choi, V. Wang, N. J. Wang, J. Ngai, J. G. Conboy, B. Parvin, H. S. Feiler, T. P. Speed, J. W. Gray, D. A. Levine, N. D. Socci, Y. Liang, B. S. Taylor, N. Schultz, L. Borsu, A. E. Lash, C. Brennan, A. Viale, C. Sander, M. Ladanyi, K. A. Hoadley, S. Meng, Y. Du, Y. Shi, L. Li, Y. J. Turman, D. Zang, E. B. Helms, S. Balu, X. Zhou, J. Wu, M. D. Topal, D. N. Hayes, C. M. Perou, G. Getz, D. Voet, G. Saksena, J. Zhang, H. Zhang, C. J. Wu, S. Shukla, K. Cibulskis, M. S. Lawrence, A. Sivachenko, R. Jing, R. W. Park, Y. Liu, P. J. Park, M. Noble, L. Chin, H. Carter, D. Kim, J. Samayoa, R. Karchin, P. T. Spellman, E. Purdom, P. Neuvial, H. Bengtsson, S. Durinck, J. Han, J. E. Korkola, L. M. Heiser, R. J. Cho, Z. Hu, B. Parvin, T. P. Speed, J. W. Gray, N. Schultz, E. Cerami, B. S. Taylor, A. Olshen, B. Reva, Y. Antipin, R. Shen, P. Mankoo, R. Sheridan, G. Ciriello, W. K. Chang, J. A. Bernanke, L. Borsu, D. A. Levine, M. Ladanyi, C. Sander, D. Haussler, C. C. Benz, J. M. Stuart, S. C. Benz, J. Z. Sanborn, C. J. Vaske, J. Zhu, C. Szeto, G. K. Scott, C. Yau, K. A. Hoadley, Y. Du, S. Balu, D. N. Hayes, C. M. Perou, M. D. Wilkerson, N. Zhang, R. Akbani, K. A. Baggerly, W. K. Yung, G. B. Mills, J. N. Weinstein, R. Penny, T. Shelton, D. Grimm, M. Hatfield, S. Morris, P. Yena, P. Rhodes, M. Sherman, J. Paulauskis, S. Millis, A. Kahn, J. M. Greene, The Cancer Genome Atlas Research Network, (Participants are arranged by area of contribution and then by institution.), Disease working group and tissue source sites, Genome sequencing centres: Baylor College of Medicine, Broad Institute, Washington University in St Louis, Cancer genome characterization centres: Broad Institute/Dana-Farber Cancer Institute, Harvard Medical School, HudsonAlpha Institute/Stanford University, University of Southern California/Johns Hopkins University, Lawrence Berkeley National Laboratory, Memorial Sloan-Kettering Cancer Center, University of North Carolina at Chapel Hill, Genome data analysis centres: Broad Institute, Johns Hopkins University, University of California Santa Cruz/Buck Institute, The University of Texas MD Anderson Cancer Center, Biospecimen core resource, and Data coordination centre. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, June 2011. Publisher: Nature Publishing Group.

[7] S. Besson, R. Leigh, M. Linkert, C. Allan, J.-M. Burel, M. Carroll, D. Gault, R. Gozim, S. Li, D. Lindner, J. Moore, W. Moore, P. Walczysko, F. Wong, and J. R. Swedlow. Bringing Open Data to Whole Slide Imaging. *Digital Pathology: 15th European congress, ECDP 2019, Warwick, UK, April 10-13, 2019: proceedings. European Congress on Digital Pathology (15th: 2019: Warwick, England)*, 2019:3–10, Apr. 2019.

[8] M. Bortolomeazzi, L. Montorsi, D. Temelkovski, M. R. Keddar, A. Acha-Sagredo, M. J. Pitcher, G. Basso, L. Laghi, M. Rodriguez-Justo, J. Spencer, and F. D. Ciccarelli. A SIMPLI (Single-cell Identification from MultiPLexed Images) approach for spatially-resolved tissue phenotyping at single-cell resolution. *Nature Communications*, 13(1):781, Feb. 2022. Publisher: Nature Publishing Group.

[9] R. J. Chen, T. Ding, M. Y. Lu, D. F. K. Williamson, G. Jaume, A. H. Song, B. Chen, A. Zhang, D. Shao, M. Shaban, M. Williams, L. Oldenburg, L. L. Weishaupt, J. J. Wang, A. Vaidya, L. P. Le, G. Gerber, S. Sahai, W. Williams, and F. Mahmood. Towards a general-purpose foundation model for computational pathology. *Nature Medicine*, 30(3):850–862, Mar. 2024. Publisher: Nature Publishing Group.

[10] M. Cummings, O. Nicolais, and M. Shahin. Surgery in Advanced Ovary Cancer: Primary versus Interval Cytoreduction. *Diagnostics*, 12(4):988, Apr. 2022.

[11] A. Dhillon, A. Singh, and V. K. Bhalla. A Systematic Review on Biomarker Identification for Cancer Diagnosis and Prognosis in Multi-omics: From Computational Needs to Machine Learning and Deep Learning. *Archives of Computational Methods in Engineering*, 30(2):917–949, Mar. 2023.

[12] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale, June 2021. arXiv:2010.11929 [cs].

[13] A. Echle, N. T. Rindtorff, T. J. Brinker, T. Luedde, A. T. Pearson, and J. N. Kather. Deep learning in cancer pathology: a new generation of clinical biomarkers. *British Journal of Cancer*, 124(4):686–696, Feb. 2021. Publisher: Nature Publishing Group.

[14] M. Gadermayr and M. Tschuchnig. Multiple instance learning for digital pathology: A review of the state-of-the-art, limitations & future potential. *Computerized Medical Imaging and Graphics*, 112:102337, Mar. 2024.

[15] A. Gillioz, J. Casas, E. Mugellini, and O. A. Khaled. Overview of the Transformer-based Models for NLP Tasks. In *2020 15th Conference on Computer Science and Information Systems (FedCSIS)*, pages 179–183, Sept. 2020.

[16] W. Grace-Jones. 25 - Tissue Microarray. In J. D. Bancroft and M. Gamble, editors, *Theory and Practice of Histological Techniques (Sixth Edition)*, pages 527–535. Churchill Livingstone, Edinburgh, Jan. 2008.

[17] M.-H. Guo, T.-X. Xu, J.-J. Liu, Z.-N. Liu, P.-T. Jiang, T.-J. Mu, S.-H. Zhang, R. R. Martin, M.-M. Cheng, and S.-M. Hu. Attention Mechanisms in Computer Vision: A Survey, Nov. 2021. arXiv:2111.07624.

[18] K. He, X. Zhang, S. Ren, and J. Sun. Deep Residual Learning for Image Recognition, Dec. 2015. arXiv:1512.03385 [cs].

[19] K. He, X. Zhang, S. Ren, and J. Sun. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. pages 1026–1034, 2015.

[20] S. Hochreiter. The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 06(02):107–116, Apr. 1998. Publisher: World Scientific Publishing Co.

[21] M. S. Hossain, G. M. Shahriar, M. M. M. Syeed, M. F. Uddin, M. Hasan, S. Shivam, and S. Advani. Region of interest (ROI) selection using vision transformer for automatic analysis using whole slide images. *Scientific Reports*, 13(1):11314, July 2023. Publisher: Nature Publishing Group.

[22] M. Ilse, J. M. Tomczak, and M. Welling. Attention-based Deep Multiple Instance Learning, June 2018. arXiv:1802.04712 [cs, stat].

[23] Jopo. jopo666/HistoPrep, Apr. 2024. original-date: 2020-12-02T13:39:00Z.

[24] D. P. Kingma and M. Welling. An Introduction to Variational Autoencoders. *Foundations and TrendsÂ® in Machine Learning*, 12(4):307–392, 2019. arXiv:1906.02691 [cs, stat].

[25] I.-M. Launonen, N. LyytikÃ€inen, J. Casado, E. A. Anttila, A. SzabÃ^3, $U. - M.Haltia, C. A.Jacobson, J. R.Lin, Z. Maliga, B. E.Howitt, K. C.Strickland, S. Santagata, K. El$ $celltumor - immunemicroenvironmentofBRCA1/2mutatedhigh - gradeserousovariancancer. NatureCommunications, 13(1) : 835, Feb.2022. Publisher : NaturePublishingGroup.$

[26] A. R. Laury, S. Zheng, N. Aho, R. Fallegger, S. HÃ€nninen, J. Saez-Rodriguez, J. Tanevski, O. Youssef, J. Tang, and O. M. CarpÃ©n. Opening the Black Box: Spatial Transcriptomics and the Relevance of Artificial IntelligenceâDetected Prognostic Regions in High-Grade Serous Carcinoma. *Modern Pathology*, 37(7), July 2024. Publisher: Elsevier.

[27] M. Y. Lee, J. S. Bedia, S. S. Bhate, G. L. Barlow, D. Phillips, W. J. Fantl, G. P. Nolan, and C. M. SchÃŒrch. CellSeg: a robust, pre-trained nucleus segmentation and pixel quantification software for highly multiplexed fluorescence images. *BMC Bioinformatics*, 23(1):46, Jan. 2022.

[28] S. Lheureux, C. Gourley, I. Vergote, and A. M. Oza. Epithelial ovarian cancer. *The Lancet*, 393(10177):1240–1253, Mar. 2019.

[29] J.-R. Lin, B. Izar, S. Wang, C. Yapp, S. Mei, P. M. Shah, S. Santagata, and P. K. Sorger. Highly multiplexed immunofluorescence imaging of human tissues and tumors using t-CyCIF and conventional optical microscopes. *eLife*, 7:e31657, July 2018. Publisher: eLife Sciences Publications, Ltd.

[30] Z. Lin, M. Feng, C. N. d. Santos, M. Yu, B. Xiang, B. Zhou, and Y. Bengio. A Structured Self-attentive Sentence Embedding, Mar. 2017. arXiv:1703.03130.

[31] L. v. d. Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[32] M. K. Mayekar and T. G. Bivona. Current Landscape of Targeted Therapy in Lung Cancer. *Clinical Pharmacology and Therapeutics*, 102(5):757–764, Nov. 2017.

[33] R. Me, R. J, J. B, G. Rp, B. Ss, R.-J. M, K. Ll, S. A, W. Hl, T. Lh, T. G, R. F, V. T. Bj, P. J, S. I, K. G, K. M, and A. Ra. High Interobserver Variability Among Pathologists Using Combined Positive Score to Evaluate PD-L1 Expression in Gastric, Gastroesophageal Junction, and Esophageal Adenocarcinoma. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 36(5), May 2023. Publisher: Mod Pathol.

[34] P. Meer. ROBUST TECHNIQUES FOR COMPUTER VISION.

[35] D. B. Nagarkar, E. Mercan, D. L. Weaver, T. T. BrunyÃ©, P. A. Carney, M. H. Rendi, A. H. Beck, P. Frederick, L. G. Shapiro, and J. G. Elmore. Region of Interest Identification and Diagnostic Agreement in Breast Pathology. *Modern pathology : an official journal of the United States and Canadian Academy of Pathology, Inc*, 29(9):1004–1011, Sept. 2016.

[36] Y. Naito and T. Urasaki. Precision medicine in breast cancer. *Chinese Clinical Oncology*, 7(3):29, June 2018.

[37] S. Narod. Can advanced-stage ovarian cancer be cured? *Nature Reviews Clinical Oncology*, 13(4):255–261, Apr. 2016. Publisher: Nature Publishing Group.

[38] M. K. K. Niazi, A. V. Parwani, and M. N. Gurcan. Digital pathology and artificial intelligence. *The Lancet Oncology*, 20(5):e253–e261, May 2019. Publisher: Elsevier.

[39] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. KÃ¶pf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. PyTorch: An Imperative Style, High-Performance Deep Learning Library, Dec. 2019. arXiv:1912.01703 [cs, stat].

[40] F. Perez-Villatoro, J. Oikkonen, J. Casado, A. Chernenko, D. C. Gulhan, M. Tumiati, Y. Li, K. Lavikka, S. Hietanen, J. Hynninen, U.-M. Haltia, J. S. Tyrmi, H. Laivuori, P. A. Konstantinopoulos, S. Hautaniemi, L. Kauppi, and A. FÃ€rkkilÃ€. Optimized detection of homologous recombination deficiency improves the prediction of clinical outcomes in cancer. *npj Precision Oncology*, 6(1):1–13, Dec. 2022. Publisher: Nature Publishing Group.

[41] S. S. Raab, D. M. Grzybicki, J. E. Janosky, R. J. Zarbo, F. A. Meier, C. Jensen, and S. J. Geyer. Clinical impact and frequency of anatomic pathology errors in cancer diagnoses. *Cancer*, 104(10):2205–2213, Nov. 2005.

[42] J. D. Seidman, R. Vang, B. M. Ronnett, A. Yemelyanova, and J. A. Cosin. Distribution and case-fatality ratios by cell-type for ovarian carcinomas: a 22-year series of 562 patients with uniform current histological classification. *Gynecologic Oncology*, 136(2):336–340, Feb. 2015.

[43] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *International Journal of Computer Vision*, 128(2):336–359, Oct. 2020. arXiv: 1610.02391 Publisher: Springer.

[44] C. Swanton, E. Bernard, C. Abbosh, F. AndrÃ©, J. Auwerx, A. Balmain, D. Bar-Sagi, R. Bernards, S. Bullman, J. DeGregori, C. Elliott, A. Erez, G. Evan, M. A. Febbraio, A. Hidalgo, M. Jamal-Hanjani, J. A. Joyce, M. Kaiser, K. Lamia, J. W. Locasale, S. Loi, I. Malanchi, M. Merad, K. Musgrave, K. J. Patel, S. Quezada, J. A. Wargo, A. Weeraratna, E. White, F. Winkler, J. N. Wood, K. H. Vousden,

and D. Hanahan. Embracing cancer complexity: Hallmarks of systemic disease. *Cell*, 187(7):1589–1616, Mar. 2024. Publisher: Elsevier.

[45] L. van Wagensveld, J. O. A. M. van Baal, M. Timmermans, D. Gaillard, L. Borghuis, S. B. Coffelt, E. H. Rosenberg, C. A. R. Lok, H. W. Nijman, L. F. S. Kooreman, J. Sanders, M. de Bruijn, L. F. A. Wessels, R. van der Wiel, C. Rausch, A. Broeks, R. F. P. M. Kruitwagen, M. A. van der Aa, G. S. Sonke, P. C. Schouten, K. K. Van de Vijver, and H. M. Horlings. Homologous Recombination Deficiency and Cyclin E1 Amplification Are Correlated with Immune Cell Infiltration and Survival in High-Grade Serous Ovarian Cancer. *Cancers*, 14(23):5965, Jan. 2022. Number: 23 Publisher: Multidisciplinary Digital Publishing Institute.

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention Is All You Need, Aug. 2023. arXiv:1706.03762 [cs].

[47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Å. Kaiser, and I. Polosukhin. Attention Is All You Need. *Advances in Neural Information Processing Systems*, 2017-Decem:5999–6009, June 2017. arXiv: 1706.03762 Publisher: Neural information processing systems foundation.

[48] K. E. d. Visser and J. A. Joyce. The evolving tumor microenvironment: From cancer initiation to metastatic outgrowth. *Cancer Cell*, 41(3):374–403, Mar. 2023. Publisher: Elsevier.

[49] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis. Deep Learning for Computer Vision: A Brief Review. *Computational Intelligence and Neuroscience*, 2018, 2018. Publisher: Hindawi Limited.

[50] C.-W. Wang, C.-C. Chang, M. A. Khalil, Y.-J. Lin, Y.-A. Liou, P.-C. Hsu, Y.-C. Lee, C.-H. Wang, and T.-K. Chao. Histopathological whole slide image dataset for classification of treatment effectiveness to ovarian cancer. *Scientific Data*, 9(1):25, Jan. 2022. Publisher: Nature Publishing Group.

[51] Y. Wang, A. J. Duval, M. Adli, and D. Matei. Biology-driven therapy advances in high-grade serous ovarian cancer. *The Journal of Clinical Investigation*, 134(1), Jan. 2024. Publisher: American Society for Clinical Investigation.

[52] T. L. Whiteside. The tumor microenvironment and its role in promoting tumor growth. *Oncogene*, 27(45):5904–5912, Oct. 2008. Publisher: Nature Publishing Group.

[53] N.-Y. Y. Wu, C. Fang, H.-S. Huang, J. Wang, and T.-Y. Chu. Natural history of ovarian high-grade serous carcinoma from time effects of ovulation inhibition and progesterone clearance of p53-defective lesions. *Modern Pathology*, 33(1):29–37, Jan. 2020.

[54] H. Xu, N. Usuyama, J. Bagga, S. Zhang, R. Rao, T. Naumann, C. Wong, Z. Gero, J. GonzÃ¡lez, Y. Gu, Y. Xu, M. Wei, W. Wang, S. Ma, F. Wei, J. Yang, C. Li, J. Gao, J. Rosemon, T. Bower, S. Lee, R. Weerasinghe, B. J. Wright, A. Robicsek, B. Piening, C. Bifulco, S. Wang, and H. Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630(8015):181–188, June 2024. Publisher: Nature Publishing Group.

Appendix A. NKI Antibodies list

Antibodies	**Fluorochrome**	**Company**	**Catalog Number**
CD11c	488	CST	45581S
CD207	555	R&D	AF2088
GranzymeB	647	Dako	M7235
CD163	488	Abcam	ab218293
CD57	555	BioLegend	359612
CD20	647	eBioscience	50-0202-80
CD4	488	R&D	fab8165g
CD3d	555	Abcam	ab208514
CD8a	660	eBioscience	50-0008-80
TIM3	488	CST	54669S
pSTAT1	555	CST	8183S
CD45RO	488	BioLegend	304212
FOXP3	555	eBioscience	41-477782
PD1	647	Abcam	ab201825
pTBK1	488	CST	14586S
CD68	555	CST	79594S
PD-L1	488	CST	14772S
CD15	488	BioLegend	301910
CD11b	555	Abcam	ab206616
yH2AX	647	BioLegend	613407
cPARP1	647	CST	6987S
Ki67	488	CST	11882
Vimentin	555	CST	9855
MHCII	647	Abcam	ab201347
CK7	555	Abcam	ab209601
MHCI	647	Abcam	ab199837
E-cadherin	488	CST	3199
a-SMA	555	Abcam	ab202509
CD31	647	Abcam	ab218582

Appendix B. Models experiments

MIL attention with DNA1, CK7 and Vimentin:	**F1 Score on Validation Set**
Trainable encoder	0.76
UNI encoder	0.84
Prov-Gigapath encoder	0.84

Table B.1: Model Performance on Validation Set, all TMA cores

MIL attention with DNA1, CK7 and Vimentin:	**F1 Score on Validation Set**
Trainable encoder	0.73
UNI encoder	0.85
Prov-Gigapath encoder	0.86

Table B.2: Model Performance on Validation Set, filtered TMA cores with low amount of cancer cells

MIL attention only with DNA1:	**F1 Score on Validation Set**
Trainable encoder	0.77
UNI encoder	0.78
Prov-Gigapath encoder	0.80

Table B.3: 1-channel Model Performance on Validation Set

MIL attention only with DNA1:	**F1 Score on Validation Set**
Trainable encoder	0.77
UNI encoder	0.78
Prov-Gigapath encoder	0.83

Table B.4: 1-channel Model Performance on Validation Set, filtered TMA cores with low amount of cancer cells

CD3d, Ki67, ECadherin, 5-fold cross validation	**AVG F1 Score on Validation Set**
UNI encoder MIL	0.84
Prov-Gigapath encoder MIL	0.86

Table B.5: Cross-validation with a different set of channels and MIL models

Appendix C. Identified ROIs

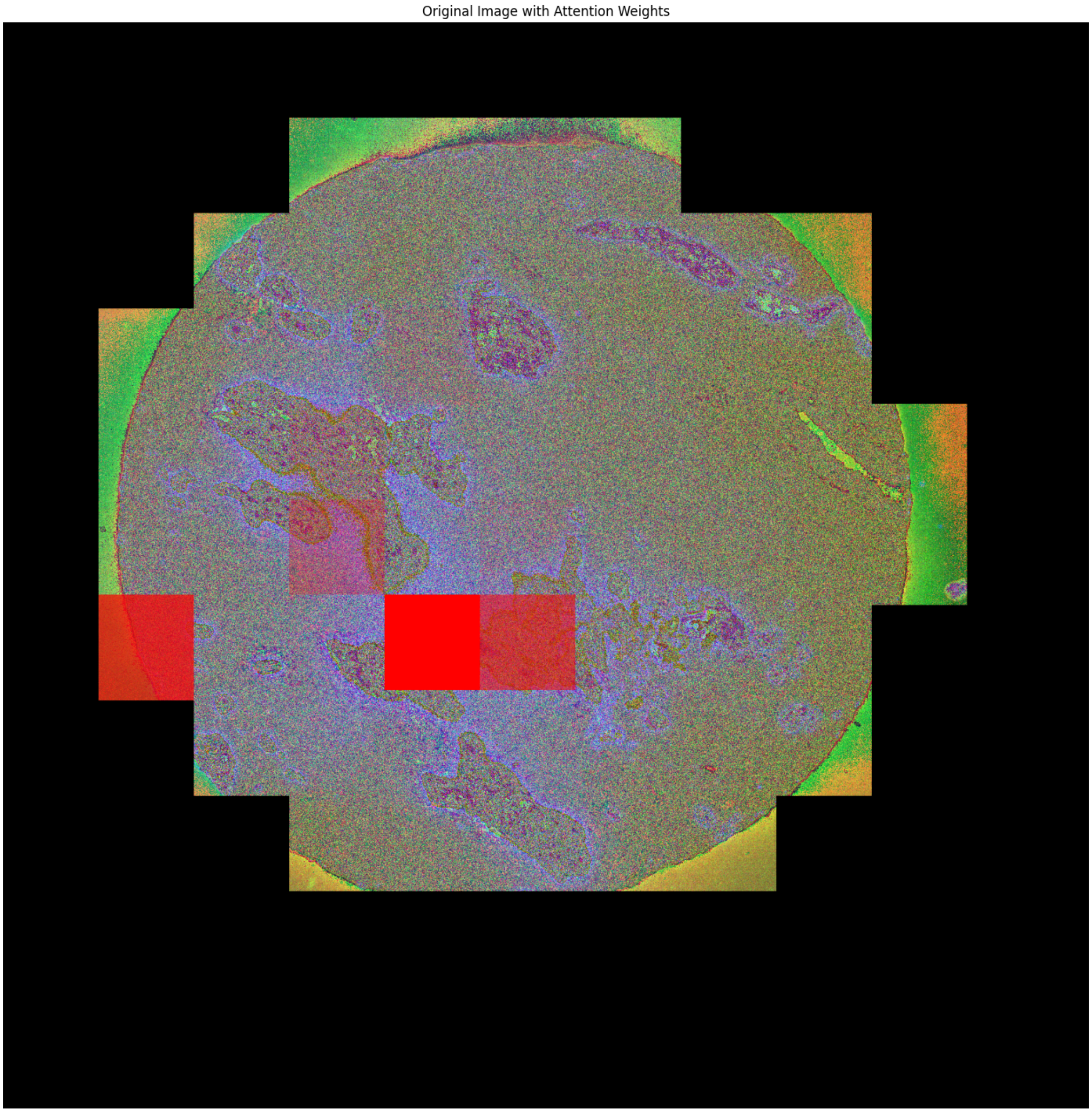


Figure C.1: ROI sections in red, by analyzing attention layer, on a TMA exposed to chemotherapy

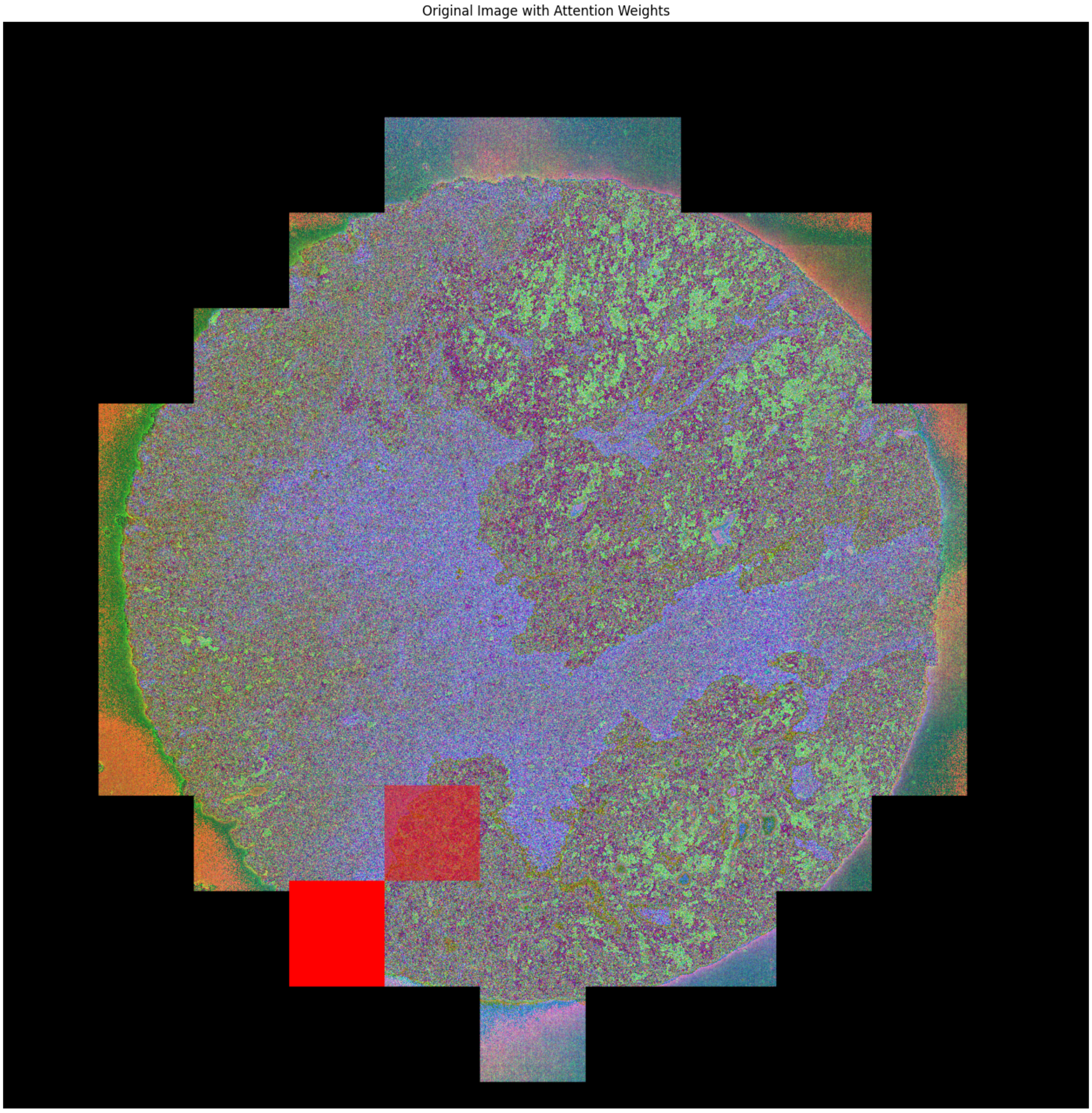


Figure C.2: ROI sections in red, by analyzing attention layer, on a TMA not exposed to chemotherapy

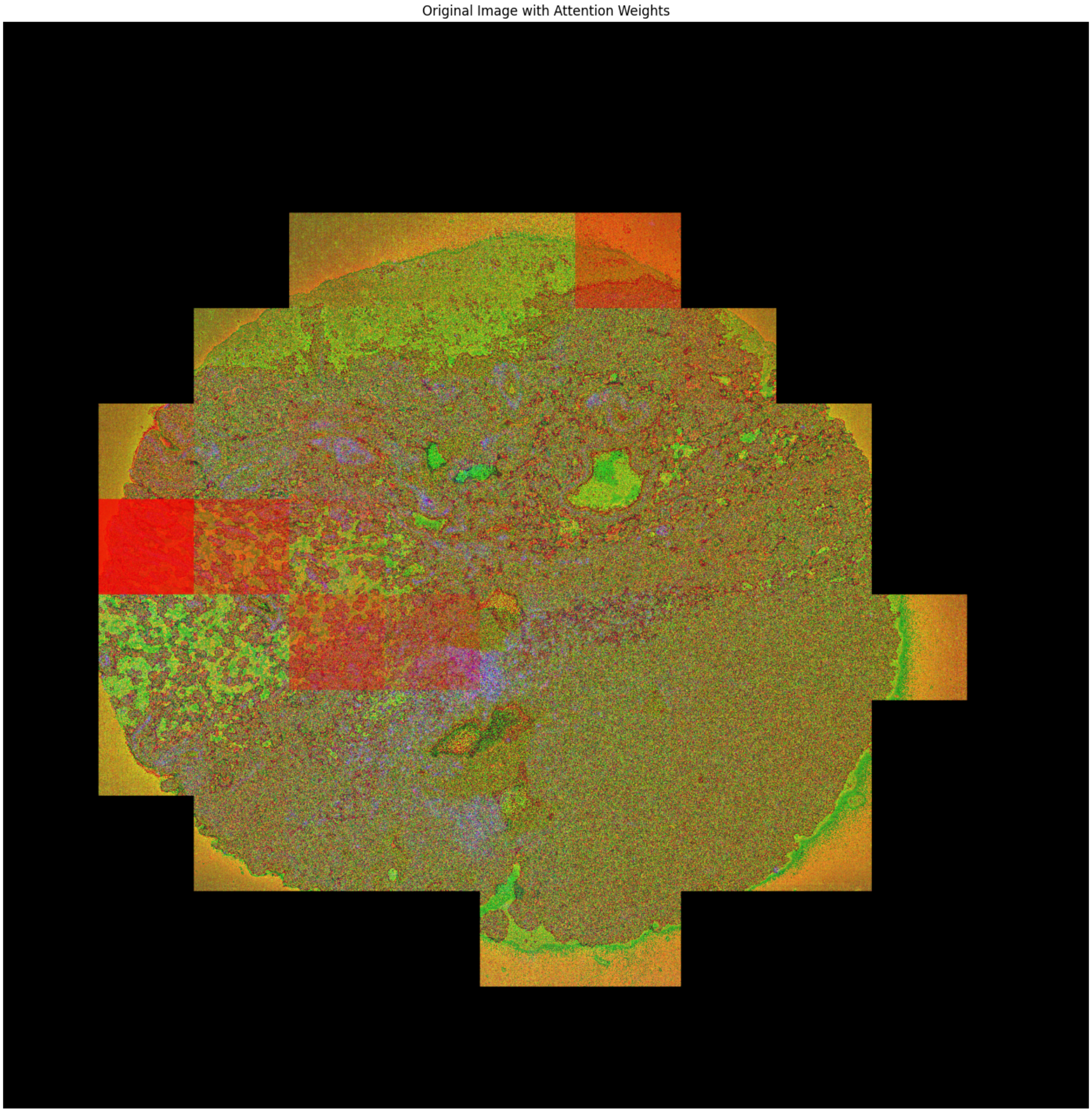


Figure C.3: ROI sections in red, by analyzing attention layer, on a TMA not exposed to chemotherapy

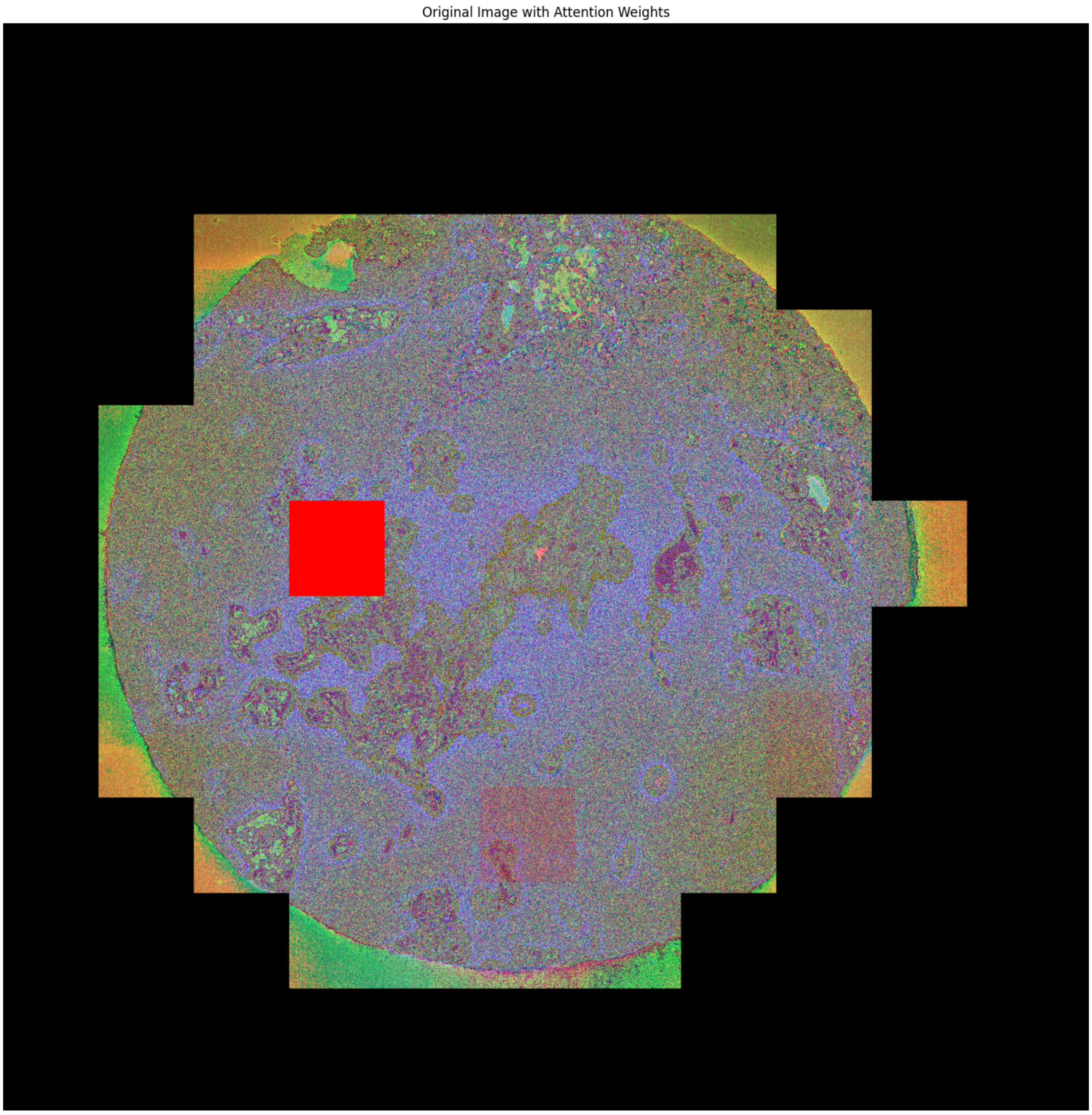


Figure C.4: ROI sections in red, by analyzing attention layer, on a TMA exposed to chemotherapy but classified by the model as not

Appendix D. Signal variation in t-CycIF data

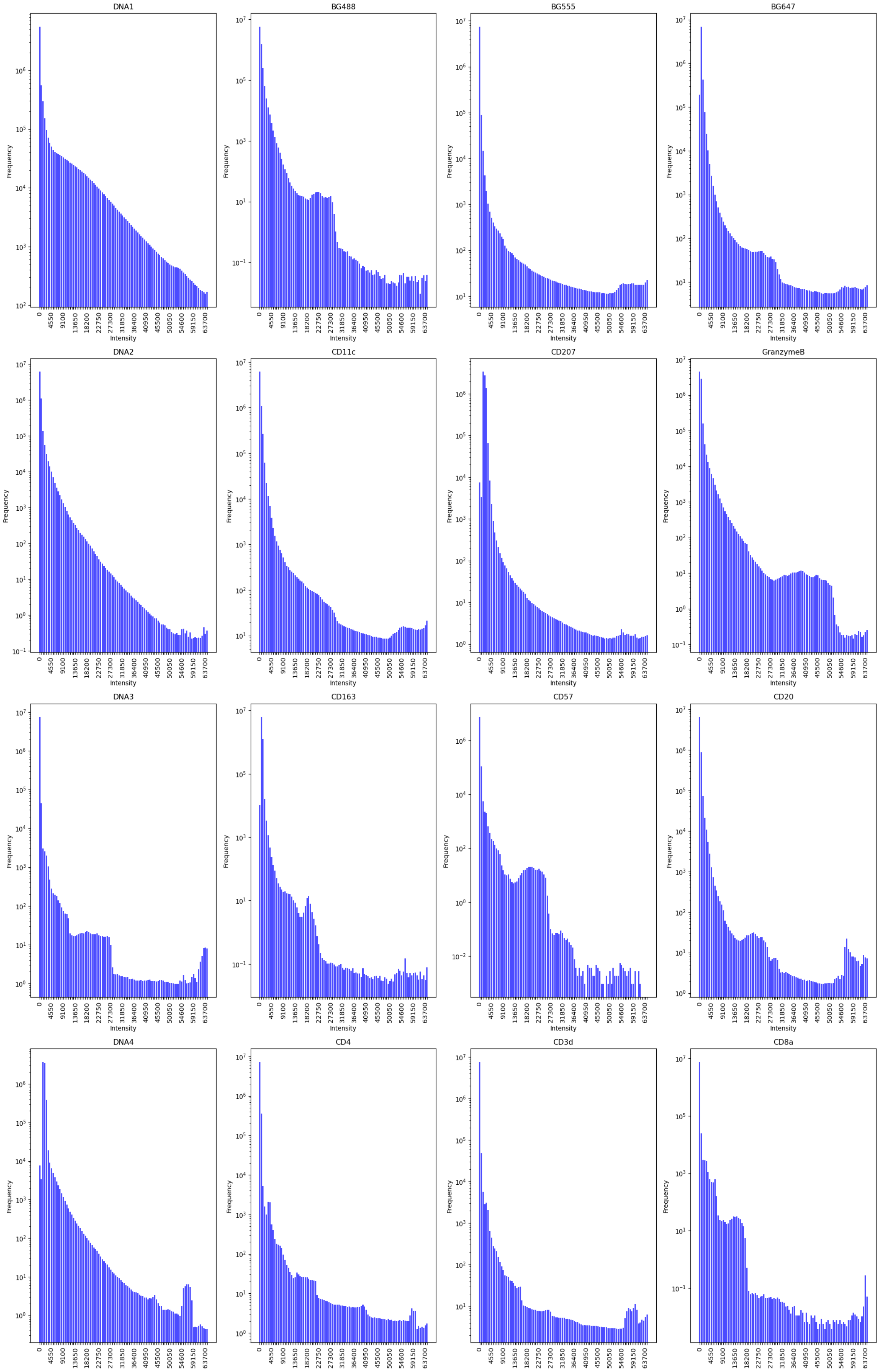


Figure D.1: Part 1 NKI cores antibodies signal.

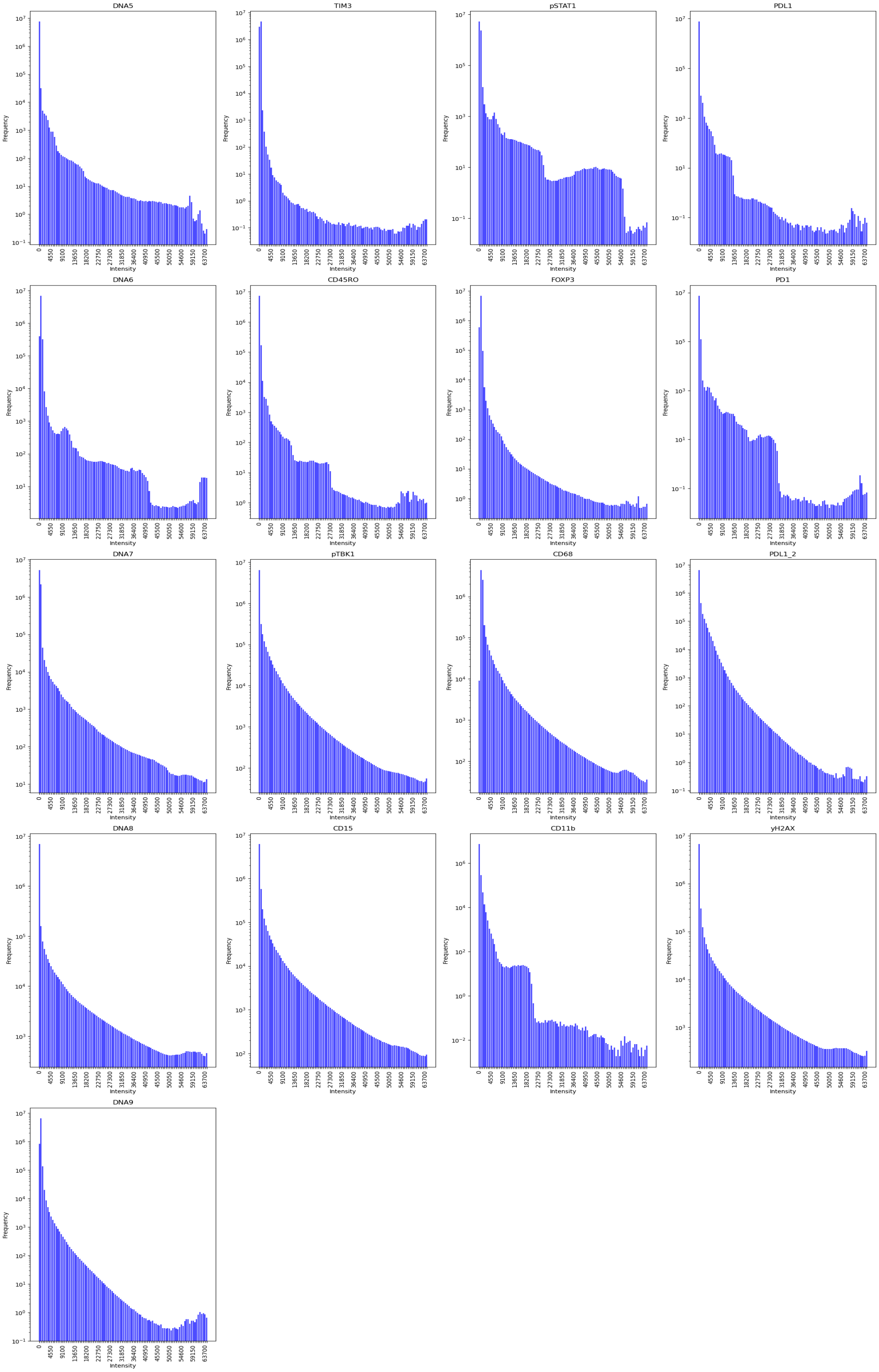


Figure D.2: Part 2 NKI cores antibodies signal.