universität wien

MASTERARBEIT / MASTER'S THESIS

Titel der Masterarbeit / Title of the Master's Thesis

„Characteristics and cell type composition of anchor-based functional marker neighbourhoods in colorectal cancer-patients“

verfasst von / submitted by

Jonathan Christ, B.Sc.

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Science (MSc)

Wien, 2022 / Vienna 2022

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	UA 066 910
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Computational Science
Betreut von / Supervisor:	Univ.-Prof. Mag. Dr. Thomas Rattei

Abstract (Deutsch)

Die technischen Fortschritte bei den In-situ-Proteom-Bildgebungsverfahren ermöglichen eine umfassende Analyse der Zelltypzusammensetzung der Mikroumgebung von Geweben.

In dieser Arbeit sollen schwach überwachte Lerntechniken eingesetzt und verbessert werden, um Merkmale der Zelltypzusammensetzung zu entdecken, die mit funktionellen Phänotypen zusammenhängen. Beispielsweise die proliferative Aktivität von Tumorgewebe, das die Nische bestimmter tumorinfiltrierender Immunzellen umgibt. Mithilfe von multiplexen In-situ-Proteom Bildern, die 56 Proteinmarker enthalten, werden Gewebsproben von Darmkrebs analysiert, um den Zusammenhang zwischen der lokalen Immunzellzusammensetzung und der funktionellen Aktivität zu untersuchen. Dies wird mit Hilfe eines Regression Modells erzielt, das maschinelles Lernen verwendet. Die sich daraus ergebenden methodischen Fortschritte könnten es ermöglichen, die Zusammenhänge zwischen der Zusammensetzung der Zellen in der Mikroumgebung und den funktionellen Merkmalen in einem allgemeinen Sinne zu bewerten.

Abstract (English)

The technical advancements of in situ proteomic imaging techniques allow for a comprehensive analysis of the cell type composition of tissue microenvironments.

This study aims to utilise and improve weakly supervised learning techniques to discover characteristics of cell type composition related to functional phenotypes, such as the proliferative activity of tumour tissue surrounding the niche of certain tumour-infiltrating immune cells. The use of multiplexed in situ proteomic images, containing 56 protein marker labels from colorectal cancer tissue slides, will be analysed to reveal the association between local immune cell composition and functional activity. This will be achieved by using a regression model, implemented in an adapted convolutional neural network. The resulting methodological advances could make it possible to evaluate these links between the composition of cell types in the microenvironment and functional characteristics in a general sense.

Acknowledgments

Frist and foremost, I would like to show my gratitude to my supervisors Prof. Dr. Manfred Claassen and Prof. Dr. Thomas Rattei for their support and guidance during this process as well as their feedback.
A special thanks to my co-supervisor Dr. Sepideh Babaei. Without her continued assistance and dedication throughout the course of this project, I could not have accomplished this thesis.

I also want to further acknowledge the University of Tübingen and my supervisor Prof. Dr. Manfred Claassen, for providing the idea for a research topic and the resources to complete this thesis, as well as for providing the opportunity to discuss and learn during our regular meetings.

Finally, I want to thank my parents Sophie and Joachim as well as my brothers Cedric and Patrick, for their encouragement and support.

Table of Contents

1. Introduction

The connection between physiological properties and spatial arrangement of cells within a microenvironment will be investigated in a data-driven method, in this thesis. The data used in this thesis was derived using co-detection by indexing (CODEX), which will be introduced in this chapter together with machine learning concepts. The foundation of the used training framework is based on neural networks, and more specifically convolutional neural networks combined with concepts of multiple instance learning. These topics will be introduced in the following. Then, the motivation and aim of this thesis will be highlighted in the last section of this chapter.

1.1 The immune tumour microenvironment in colorectal cancer

The tumour microenvironment (TME) is a complex mixture of tumour, immune, and stromal cells all interacting with each other; thus understanding the composition of the TME, specifically with respect to rare cell subtypes and at the single-cell resolution, plays a crucial role for understanding the progression of a tumour and developing therapies (Gohil et al., 2021). Several authors highlighted the importance of understanding the spatial organization of cells, to be able to connect and associate it to the underlying high-level functionality of a tissue and its molecular profiles, for instance by the use of transcriptomic data and tissue imaging technologies (Bae et al., 2021; Harold, 2005). The composition of the immune tumour microenvironment (ITME) can be linked to cancer progression and effectiveness of immunotherapies for various cancer types, such as the role of tumour infiltrating immune cells or the complex ITME organisation which may explain ways in which tumour cells evade immune responses (Jansen et al., 2019; Labani-Motlagh et al., 2020). Being able to control and change the ITME is crucial to improve prognosis of cancer patients and respective therapy approaches, including the presence of several polypeptide growth factors and their receptors in the ITME, such as the epidermal growth factor receptor (EGFR, see Table 1). These growth factors and their mechanisms are believed to play major roles in the ITME, including proliferation and differentiation pathways (Chew et al., 2012; Zhang et al., 2010). In the article by Schürch et al. (2020), the authors looked at the microenvironment of colorectal cancer patients to understand the connection between the ITME and survival. The team also tried to explain how various ITME compositions and functions could interact with one another, carry out specific tasks, and contribute to the overall progression. One such investigation, attempted to link enrichment of certain cell types within a neighbourhood to the

overall survival of a patient. For this objective, the authors compared the neighbourhoods of two distinct patient groups: “Crohn’s-like reaction” (CLR) and the diffuse inflammatory infiltration group (DII). The CLR group has a high formation of tertiary lymphoid structures (TLS) along the invasive tumour front, while the DII group does not show formation. These two groups have been matched in terms of their characteristics, such as age and gender but differ in their TLS expression, and come from two opposing TLS expression spectra, which have been demonstrated to have therapeutic significance for several cancer types. Tertiary lymphoid structures (TLS) occur at sites of infection or persistent inflammation and have been observed in autoimmunity, transplant rejection and according to more recent studies, malignancy. The clinical relevance of TLS is believed to range from detrimental to preventive (Colbeck et al., 2017). In the elimination and equilibrium phase, when the immune system eliminates tumour cells and regulates the overall progression of the tumour, it has been shown that high TLS density is favourable for a patient’s survival and affects general progression in patients with colorectal cancer patients (Trajkovski et al., 2018).

1.2 Co-detection by indexing; A spatial proteomic imaging technique

The dataset used in this thesis, produced by Schürch et al. (2020), is based on the co-detection by indexing (CODEX) technology. A crucial function in both basic research and clinical treatment is the ability of antibodies to attach to recognized proteins and to help with identification and quantification of cells and cell phenotypes. For pathologists and scientists, immunohistochemistry and immunofluorescence imaging are important techniques for determining the organization of cellular and subcellular proteins. CODEX is a multiplexed tissue visualization method and a platform for fluorescence microscopy. CODEX detects utilized DNA-conjugated antibodies and fluorescently labelled, complementary DNA probes, which are inserted and removed in a cyclic fashion, in order to visualise up to 60 markers in situ concurrently. CODEX allows research into the architecture and cellular structure and gives insights into regulators of pathological conditions such as infections, tumour immunology, and immunotherapeutic targets (Phillips et al., 2021). Although these methods are of great clinical importance in diagnostics, they are limited in the number of protein markers that can be evaluated in a particular tissue segment. A variety of multiplexed tissue imaging technologies exist that significantly improve the capacity to detect and profile complex biological systems with respect to spatial information, at a single cell resolution.

Other methods, in addition to CODEX, use Raman scattering, Trapped Ion-Mobility Spectrometry (TIMS) or Matrix-Assisted Laser Desorption/Ionization (MALDI to create multiplex images. However, these procedures require specialised equipment and knowledge, have insufficient resolution, or have restricted multiplexing capabilities. CODEX provides a number of benefits over other multiplexed imaging methods and reduces the number of technological obstacles to multiplexed tissue imaging. Multiple inverted fluorescence optical microscopes are used by the CODEX technique, which are often accessible in research laboratories, which may greatly reduce initial infrastructure expenditures (Kennedy-Darling et al., 2021). Original CODEX used fluorescent deoxynucleotide triphosphate (dNTP) analogues and DNA polymerase primer extension1 to make antibodies visible.

A CODEX based setup can be scaled, antibody panels personalised, and the device has adequate throughput to capture tissue microarrays and samples up to 1 cm. However, it lacks a signal amplification mechanism. CODEX is suitable for the use with formalin-fixed, paraffin-embedded (FFPE) and fresh-frozen (FF) samples as well as individual cell spreads, for instance (Black et al., 2021). The data used in this thesis the data was created using the FFPE method (Schürch et al., 2020), re-designed and fully automated the CODEX FFPE workflow, to increase efficacy in FFPE tissue, among other improvements.

FFPE is a method for preserving and preparing samples that is useful for evaluation, experimental research, and drug development, for instance, by fixing a tissue sample in formaldehyde (formalin) and submerging it in paraffin, a type of wax, to preserve proteins and essential structures within the tissue (Cazzato et al., 2021).

Characterizing several cell phenotypes concurrently, across several cell types, using an assortment of proteins has been shown to be a requirement for predicting therapy results. These modalities enable only a small number of proteins to be studied at once.

Multiplexed single-cell phenotyping techniques, such as flow or mass cytometry-based methods, may simultaneously evaluate thousands of cells with more than forty antibodies. Nonetheless, the tissues must be separated in cell suspensions, resulting in the loss of spatial configurations of the studied cells and the possible omission of some cell types. Therefore, multiplexed tissue imaging technologies have been developed to study these critical spatial features alongside other cellular features. Multiple kinds of detecting techniques support multiplexed imaging strategies, such as CODEX (Black et al., 2021).

1.3 Neural networks and multiple instance learning

CellCnn, created by Arvaniti and Claassen (2017), is the primary analytical framework of this thesis. CellCnn specialises in the identification of uncommon cell populations, which play a crucial role in the overall state of eukaryotic organisms.

To achieve this objective, techniques from multiple instance learning and convolutional neural networks) have been combined. Training a network optimizes weights to fit the phenotype of the training data set, which is then correlated to the molecular profiles of the relevant subgroups of cells (Arvaniti and Claassen, 2017). Concepts related to machine learning, convolutional neural networks and multiple instance learning will be introduced in this section.

Neural Networks

Neural networks are a sub-category of machine learning, that aim to mimic the human brain by means of a set of algorithms. These algorithms attempt to recognize underlying patterns and relationships in a dataset. The main difference to regular machine learning is the way in which the algorithms learn, requiring less human intervention for the analysis of large datasets during training. As an example, an important aspect of machine learning is feature engineering, which means feature selection and extraction. Deep Learning excels at extraction features, which would otherwise be a time-consuming task, which requires in-depth knowledge of the problem, which it not always available beforehand. A neural network is made out of four general components: inputs *x*, weights *w*, a bias *b* and an output f(x). The formula could look like this:

$$\text{f(x)} = \sum_{i=1}^{m} wixi + b = w1x1 + w2x2 \ldots + wmxm + b$$

To calculate the output, an activation function is defined. For example, a sigmoid function could be used. If the output of any node is above the specified threshold, the node sends data to the next node of the next layer in the network. This process can be initiated with an arbitrary number of hidden layers, each could have their own activation function. Finally, the output of all hidden layers is combined to calculate the final output of that neural network. Typically, a real-world problem will have many hidden layers and will be non-linear. In contrast to regular regression, the effect that each weight has on the overall problem of a neural network is much greater. In regression a single weight can be changed without changing the output in other functions, this is not possible for a neural network (Sharquay, 2020).

Deep learning uses a neural network with many hidden layers, according to some definitions a neural network with at least three hidden layers could be considered a deep learning approach. Figure 1 below depicts a simple neural network.

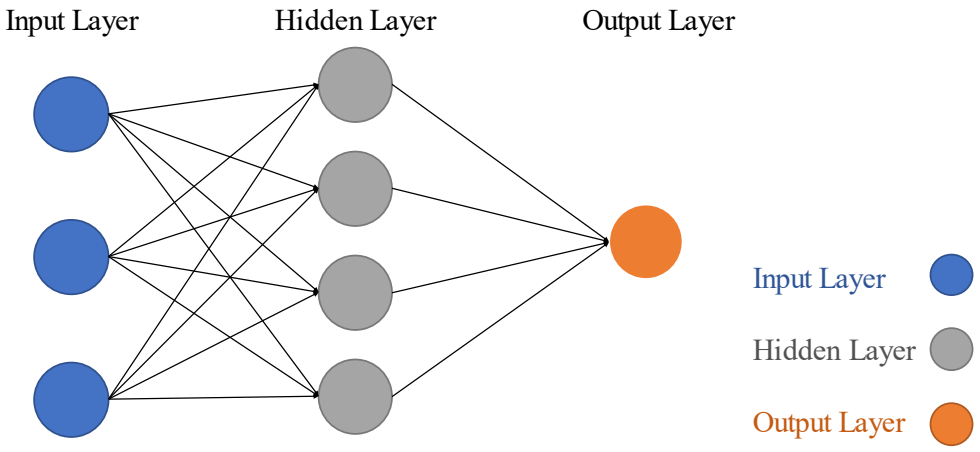


Figure 1: Shows a simple Neural Network, with three input nodes connected to four hidden layers that connect to one output layer.

There are two general directions to train your neural network. Feed-forward, going from the input to output in only one direction and backpropagation, which goes from the output to the input allowing for adjustment to the fitting algorithm by using an associated error with each neuron. There are three types of neural networks, convolutional neural networks (CNN), recurrent neural networks (RNN), and artificial neural networks (ANN). ANN are a group of multiple perceptrons or nodes (a single perceptron could be seen as a logistic regression).

A disadvantage of an ANN is that it deals with the vanishing and exploding gradient problem during backpropagation, in which the depth of the network corresponds to a loss of training capability as the size of the training set increases and as the gradient decreases (Tarnate, 2020). ANN also cannot learn sequential data. When a regular ANN has a looping constraint on the hidden layers, it becomes a recurrent neural network (RNN), which due to their cyclic graphs allow for the capture of sequential data. RNN are mostly used for audio, text data, and time series data. The loop means that each of the input data points is dependent not only on the output of the current node but also all the previous nodes, and thus allows for the transfer of data across time steps (Hochreiter, 1998; Lipton, 2015). The neural network used in this thesis is a modified convolutional neural network (CNN), which excels at image and video

processing. The main components of CNN's are filters, also called kernels, which extract the relevant features, without explicit reference.

Convolutional neural networks can learn the spatial features of an image, thus the arrangement of pixels and the underlying relation between them in the image. A single filter can be used to produce a feature map, by being applied to different parts of the input data image (Namatēvs, 2017).

Convolutional neural networks

The backbone of CellCnn, the framework used in this thesis, is convolutional neural networks; hence they will be introduced in this section in more detail. The inspiration for convolutional neural networks comes from the Visual Cortex, where individual neurons only react to specific areas of our visual input space and, by overlapping collections of such areas, obtain the entire vision (Lindsay, 2021). A convolution is a linear process that, like a conventional neural network, multiplies a set of weights and adds a bias with the input, carried out by a convolutional layer. A feature map shows the positions and intensity of an observed feature in an input, such as an image or video, which is produced by repeatedly applying the same filter to an input, this process is depicted in Figure 2.

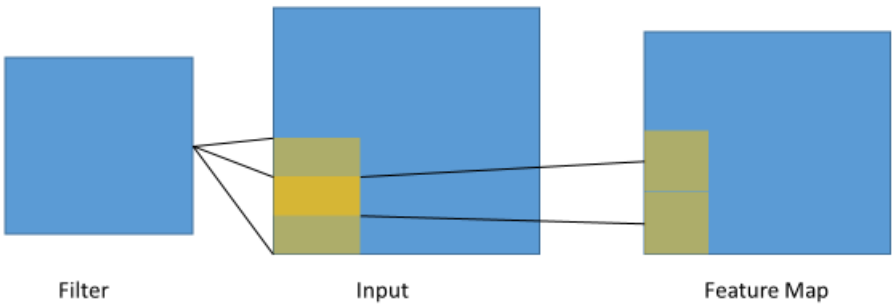


Figure 2: The filter repeatedly overlaps the input image and by doing so creates the feature map.

A CNN can automatically learn from several filters tailored to a training data set in parallel, while adhering to the limitations of a particular predictive modelling issue, like image classification. Those resulting highly specific features can then be detected anywhere in the input data. Although they can also be applied to different dimensionality data, CNN are typically used with two-dimensional image data (Namatvs, 2017). A Filter or kernel traverses the image to determine whether a certain feature is present or not. Multiplication is performed between an array of input data, which is a two-dimensional array of weights, and a

filter or kernel for two-dimensional input. The size of the filter, which also determines the size of the receptive field, is typically a 3x3 matrix, although the size may vary. A feature map, activation map, or convolved feature is the result of the input and filter's dot product sequence. The pooling then shrinks the feature map. Multiplying an input patch by the size of the filters and the filter itself produces an element-wise dot product. The result is then summed up and the bias is added, if applicable, then becomes a single output value. An activation function, which maps the input and output to one another, can be used to further control the activation of a node in the network. The filter is often smaller than the input data because it enables the same filter, which are a set of weights, to be multiplied by the input array repeatedly at various locations on the input. The filter is systematically applied to each overlapping area, with the size of the filter in the data, and may be implemented to discover a certain type of feature in the input to continually locate that feature anywhere in the data (Alzubaidi et al., 2021).

Translation invariance is the term used to describe this ability, which refers to the ability to recognize a certain distinctive feature in the data, even after applying geometric, spatial translational operations. A more formal way to define the discrete convolutional operation is as follows.

$$c(t) = (f * g)(t) := \int_{-\infty}^{\infty} f(\tau) g(t-\tau) g(\tau)' \quad (2.1)$$

Where f and g are the functions to be convoluted, c describes the resulting function, t is a variable of real numbers, $g(\tau)$ is the convolution of $f(t)$ and $g(\tau)'$ its derivative (Biscione & Bowers, 2020). Figure 3 below depicts a simple CNN concept.

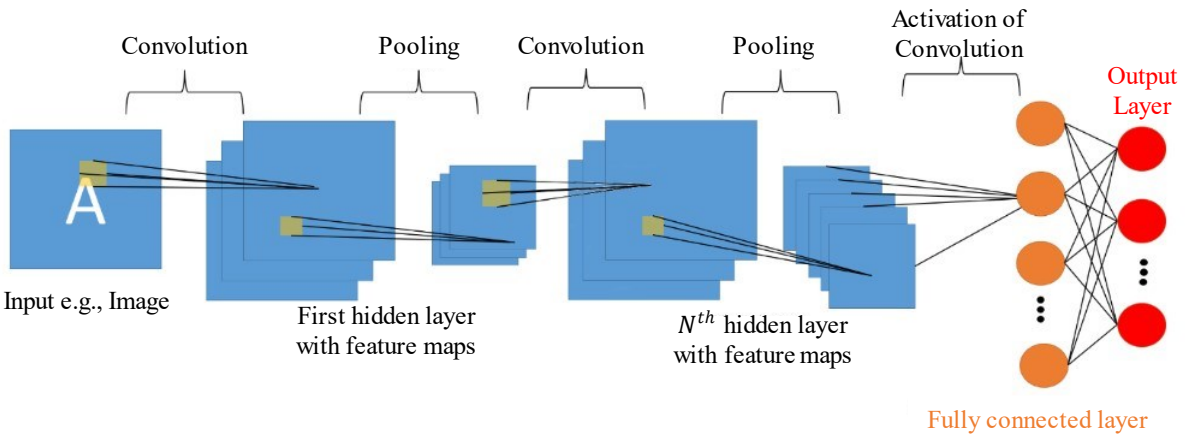


Figure 3: Depiction of a simple CNN concept. The first layer of a CNN is the convolutional layer and can be followed by either a pooling or further convolutional layer. Before the last layer, there is typically a classification layer that uses an activation function, which is then connected to the final fully connected layer.

Typically, low-level characteristics like edges, colour, gradient direction, etc. are captured by the first convolutional layer. With more layers, the architecture adjusts to high-level characteristics as well, giving the network a way to comprehend the dataset much like a human would. The procedure yields two different types of results: one in which the dimensionality of the convoluted feature matrix is decreased as compared to the input using valid padding, and the other in which it is either increased or stays the same using the same padding. However, the size of the output can also be increased using full padding, where the output is expanded by adding zeros to the border of the input. As can be seen in Figure 3, there is also pooling to consider, which decreases the spatial dimensionality of the convoluted features by using dimensionality reduction while maintaining the translational invariance. Dimensionality reduction refers to a number of strategies to reduce the dimension of the problem, for example, by using feature selection methods, matrix factorizations and eigenvalue decompositions.

The two types of pooling are average and max pooling. Average pooling returns the average of all values considered by the filter or kernel, while max pooling returns the maximum value (O'Shea & Nash, 2015).

Multiple Instance Learning

Multiple Instance Learning is a kind of weakly supervised learning technique, where in the case of a binary classification issue, training data is organised in bags, each of which includes a collection of examples and has a single label. Although they are unknown during training, each instance in a bag has a unique label. Suppose that there are several instances; a bag is said to be negative if all of its instances are negative but is considered positive if it contains at least one positive instance. Instance bags: X = {x1, x2, x3, …., x_M}, where X is a bag with M instances and a label Y, where each feature xi directly relates to M individual labels y_i, which are part of Y $\boldsymbol{\epsilon}$ {0,1}. For image-oriented problems, such as object identification in an image, instead of bags there are instances which are to be identified. The task is to check for the presence of that instance within the image. Since training is carried out using data organised in sets, instance classification differs from bag classification in that the goal is to categorise each instance separately. Misclassifying an instance when the aim is bag classification does not always result in a loss at the bag level. Numerous approaches that have been suggested for classifying bags do work in an instance space, and thus often cannot achieve instance classification. For regression tasks, instead of assigning class labels, a real value is assigned to a bag or instance instead (Carbonneau et al., 2016).

1.4 Motivation and aim of the thesis

This thesis will build on the foundation of Schürch et al. (2020) who investigated the immune tumour microenvironment in terms of their cellular neighbourhoods and their cell type composition. This thesis will aim to expand on this premise. Analysing how these cellular neighbourhoods interact with each other can provide new insights into the ways cancer affects the structure, function, and communication within the immune tumour microenvironment. Some tumour types may avoid immune responses by excluding T cells or overexpressing cell types such as macrophages and regulatory T cells. Clarifying the function of the microenvironment and its role in tumour progression has become a major area of study but investigating the correlation of spatial organization and tumour immune response with respect to the communication within an ITME has not been extensively studied. Assessing tissues merely in terms of their interacting cell types may not be reliable enough by itself to accurately portray tissue function (Schürch et al., 2020).

In order to view cell types and the cellular neighbourhoods together, CellCnn (Arvaniti & Claassen, 2017) has been chosen as the framework of this thesis. CellCnn has initially been used in a classification setting and will be used, as part of this work, for the first time with a weakly supervised regression setting to allow for the cell type composition and the cellular neighbourhood to be viewed together and analyse the immune tumour microenvironment across several colorectal cancer patients.

Unsupervised learning strategies, applied to the investigation of the TME, may only use single-cell data without further information about the phenotype or state of the disease. These strategies allow for classification of cell type compositions, but without consideration of rare cell subtypes based on conditions, that can be further defined by labels in supervised learning. These rare cell subtypes derived from supervised learning approaches can provide new insights into yet undiscovered relations and are currently not the subject of much research (Arvaniti & Claassen, 2017; Babaei et al., n.d., Appendix I). An ordinal regression model using CellCnn will also be implemented as an alternative approach, with the same overall goal.

The aim of this paper can be summarized to the successful application of CellCnn with a regression model, to learn the correlation between cell types and cellular neighbourhoods within the ITME in colorectal cancer patients. This will be achieved by building a neighbourhood of cells around an anchor cell type, with a specific functional marker expression. The aim is to detect enrichments for considered cell types in the proximity of an

anchor cell with a chosen functional marker expression. Different strategies aimed at training with CellCnn and optimizing the results as well as an alternative ordinal regression problem will be investigated. The resulting enrichment scores will be visualized and analysed in a biological context. By doing so, the cell type composition of the investigated cell types will be analysed and might allow general insights into the functionality and characteristics of the underlying immune tumour microenvironment.

2. Methods

This section will explain the framework, data, and procedures used in this thesis.

During this thesis, multiple different functional markers were used. The first functional marker was ki67, a proliferation marker. Cell markers and functional markers were pre-processed as stated in section 2.1.3. During training, both survival groups were considered. An anchor cell type was chosen, one around which the nearest neighbours were computed for each patient separately. 55 of the 56 markers were used as input during training, excluding the functional marker used as a label. Once the nearest neighbours were computed, the corresponding functional marker intensities for each of the cells in the neighbourhood were considered to compute the label. Here, different computation modes have been implemented, which will be explained in the results section, and then compared to each other to establish a trend that could capture the underlying pattern with the most accuracy. As further illustrated in the results section, further functional markers had to be considered: EGFR, ICOS, BETA-CATININ and BCL-2, which will all be further explained in Table 1. Additional strategies such as binning and ordinal regression will also be explored. The ordinal model used the binned intensity data, which is also ordinally sorted, to train the model instead of the regular intensities. An enrichment score was calculated to analyse the biological relevance of the selected cells. This will be briefly stated in the Methods section along with other relevant theories for the downstream analysis, but this will be further explained in the results section.

2.1 The dataset

In the following section, concepts related to the data set will be explained. First, information about the structure on the dataset itself will be provided, and then, the final two subsections will explain which marker labels have been chosen for this thesis and how the data were pre-processed prior to training.

2.1.1 Data structure

The data used in this thesis, to train CellCnn, stems from the team of Schürch et al. (2020). In their study, they used co-detection by indexing (CODEX, see section 1.2) for paraffin-embedded tissue microarrays, which enabled the parallel assessment of 140 FFPE tissue regions in 35 colorectal cancer patients. Nine unique cellular neighbourhoods and 28 cell types were observed, each containing 56 annotated functional markers, resulting in a total of 258,385 cells. These cells have been further grouped into tissue microarray regions called TMA regions.

The 35 patients come from the 2 different patient groups CLR and DII, mentioned in the introduction, from two different spectra of TLS formation. The 56 markers consist of 20 immune markers, 21 functional markers and 15 auxiliary markers. For the feature labels of this thesis, only functional markers were considered, with the input containing the remaining 55 markers. The spatial coordinates of every cell inside the tissue can be found by analysing such a CODEX image. The resulting CSV data structure was used for the analysis, where each row is made of one cell, with a column for the cell id, used for later downstream analysis, a survival group (DII or CLR), an annotated cell type from 28 unique cell type clusters, marker intensities of the functional markers, the position of the cells (x/y coordinates) and their corresponding TMA regions, among other values which are not relevant for this thesis. Schürch et al. have chosen TMA regions to capture cellular phenotypes that particularly characterised the immunological tumour microenvironment on the invasive tumour front (Schürch et al., 2020).

2.1.2 Functional markers used in this thesis

The following Table 1 sums up the markers used during the analysis.

Table 1: Overview of functional markers used during analysis

Functional Marker	Function Overview
Ki67	Nuclear protein that is commonly affiliated with the regulation of proliferation.
EGFR	Epidermal growth factor receptor: A transmembrane protein of the receptor tyrosine kinase family, that might be involved in angiogenesis and control of proliferation among other functions. In cancer, excess EGFR can cause rapid cell division and accelerate cancer progression of cancer (Fang & Wang, 2014).
ICOS	Inducible costimulator; Could act as a conserved immune response mediator and is mainly induced after T cell activation. ICOS can be found in the vicinity of antigen-specific T cells, cytotoxic T cells, memory- and regulatory T cells. It acts as a costimulator for T cell proliferation. (Xiao et al., 2020)
Beta-Catenin	Key component of the Wnt signalling pathway, a highly conserved evolutionary route that modulates various vital cell functions in animals that are essential for the growth and general progression of an animal's longevity (Valenta et al., 2012). Studies suggest that multiple malignancies may result from an unbalanced beta-catenin distribution (Wang et al., 2020).
BCL-2	B cell lymphoma-2; BCL-2 is a family relevant to the regulation of apoptosis. It's an important protein for regular B cell differentiation and development, and acts in anti-apoptotic processes, where it contributes to overall cell survival. However, upregulation could lead to favourable conditions for malignant B cells (Kale et al., 2018).

2.1.3 Pre-processing of Data

Before using the data for training, the data was pre-processed using the statistical methods of a logarithmic (log) offset and z-score scaling across the entire dataset, containing all patients and regions.

$$Offset = \log(1e - 3 + x)$$

$$Z - Score = \frac{xi - x}{\sigma}$$

Where σ is the standard deviation of the data (x) and xi is the i^{th} data point of x. The z-score is a normalization method that depicts how far away the data point is from the standard deviation, and thus depicts the position of each datapoint on a normal distribution.

2.2 Implementation of the Framework

This section explains concepts related to the used framework of this thesis in more detail. The focus will be CellCnn, a convolutional neural network that adapts ideas from multiple instance learning, and concepts associated with CellCnn.

2.2.1 Kolmogorov-Smirnov two-sample test

The Kolmogorov–Smirnov test is a nonparametric equality test for continuous and discontinuous probability distributions. In the case of two samples, the distribution is evaluated under the null hypothesis and is used to determine whether or not two independent samples originate from the same distribution. This test is significant because, in many cases it specifies the method that should be used to estimate the model's unknown parameters and the testing procedures that the analyst may use. It computes the maximum difference between an empirical and a hypothetical cumulative distribution. The score can be generally defined as

$$D_{m,n} = \max | F(X) - G(X)|,$$

where the observed cumulative distributions of a sample of size m or n are denoted as F and G respectively, and we define the samples as F(x) = k/n or G(x) = k/m, where k is the number of data point ≤ X, and $D_{m,n}$ the reference distribution. Scores of both groups are organised in a cumulative frequency distribution using the same intervals or categories. The cumulative distributions are then compared to assess whether both samples were chosen from the same

distribution, which is true when both distributions are reasonably close to each other and assumed false otherwise (Watada, 2010).

2.2.2 CellCnn

CellCnn is a representation learning method that employs high-dimensional single-cell data to find unusual cell subsets linked via pathology. It combines elements of multiple instance learning and convolutional neural networks as well as the advantage of large datasets in order to determine a representation for a cell population, which is then associated with a disease state. During network training, weights are optimised such that the predicted phenotypes from the network match the actual phenotypes. The membership of individual cell subsets may be attributed using filter weights, assigned during training, that correspond to the molecular profiles of relevant cell subsets (cell filter responses). If the weights of a filter and the markers of certain correspond to each other, the result would be a high cell filter response, if they don't correspond to one another, a low filter response is the result. In some instances, a cell subset selected by a filter may include more than one kind of cell, all of which are associated with the characteristic under investigation, such as a stimulus reaction. Density-based clustering is used on all of selected cells by a filter, with respect to all observable variables. A quantitative score derived from the Kolmogorov-Smirnov test, introduced in section 2.2.4, determines the distinguishing markers of each selected cell type that were computed for each marker, the difference between all available cell types and the selected cell type is then summed up, with respect to the occurring marker distribution. Measurements of individual cells correspond to imaging patches. For each cell measurement, the output of each convolutional filter is independently evaluated by the pooling layer. The activity of the convolutional layer nodes is given as weighted sums of single-cell molecular profiles. Pooling layer nodes, either evaluate the appearance or frequency of chosen cell subsets, using either pooling with the maximum or mean, referred to as max-pooling and mean-pooling respectively. The response is computed on the multi-cell input. The appearance of cells providing a high filter response is determined via max-pooling, which calculates the highest response among all cells of a multi-cell input for a particular filter. By analysing the average filter response of cells in a multi-cell input, mean-pooling determines the frequency of the subset of cells that respond strongly to a given filter. The pooling layer connects to the output layer. Regression problems use a single node while one node per class is provided for classification problems.

The network's output predicts the sample-associated trait (Arvaniti & Claassen, 2017).
The paper in Appendix I (Babaei et al., n.d.) is derived from the original work of Arvaniti & Claassen (Arvaniti & Claassen, 2017) and is featured at the end of this thesis, as it is not yet published. Figure 4 below shows the concept of CellCnn.

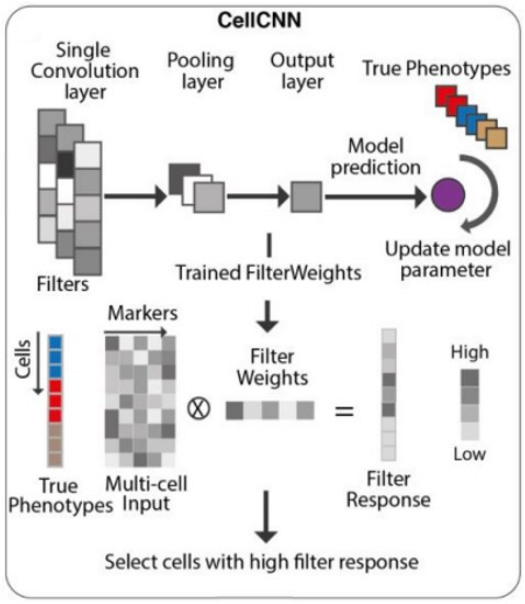


Figure 4 shows the concept of CellCnn, an adapted CNN, with a single convolutional layer and the pooling layer as described above. This Figure is a section of "Fig.1" from the paper (Babaei et al., n.d.), featured in the Appendix I.

2.2.3 Ordinal regression

Ordinal regression settings are machine learning problems, where the aim is to categorise, patterns using a natural order scale between labels, and the target variable labels indicate natural ordering. Ordinal regression is often used in medical research, age estimation, and brain-computer interfaces, all of which yielded better results with ordinal regression models than their nominal counterparts. In machine learning and data mining, learning to classify or predict numerical values using prelabelled patterns is one of the primary areas of research. Between the two primary categories of ordinal categorical variables, the terms "grouped continuous variables" and "evaluated ordered categorical variables" are often used. The first is an observable discrete representation of an underlying continuous variable. The second contains the variables for which an individual gives a grade to the ordered categorical variable. However, in both cases, the implementation of an order is significant (Gutiérrez et al., 2015). In this thesis, ordinal regression will be applied to the ordered, pre-processed intensity values of the feature vector, where the intensity values are ranked from lowest to highest. The intensity

values will first be binned together into groups, in case of 5 bins, the values will be separated into 5 ordinal values, where 1 would be the lowest and 5 the highest intensities of a given marker label. Those ordered integers will be the input for ordinal regression in this work.

2.2.4 Activation functions (TensorFlow)

The main activation function used during training is the ReLU activation function, which is often used for convolutional neural networks. Integer values of the input are converted into positive numbers. The function can be described as

$$f(x)ReLu = \max(0, x).$$

The ReLU function reduces computational cost, but can potentially lead to problems during backpropagation known as 'dying ReLU' (Filho et al., 2011). The final layer of the CellCnn is then a linear activation function, to produce the regression output.

2.2.5 K Nearest Neighbours

The multi-cell input is the set of marker expression vectors, using the marker intensities as quantity, computed in one of the ways described in 2.2.5. The set includes the k-nearest neighbour cells $[Ci]_{KM}$ $i \epsilon \{1, 2, \ldots, N\}$ where K and M are the number of cells and markers, respectively. The distance between the K cells was computed using the Euclidean distance, which is defined as:

$$d(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

where x and y are two points and x_i and y_i are the corresponding vectors (Agostino & Dardanoni, 2009).

2.3 Performance Evaluation and assessment of results

In order to assess the performance and significance of the methods used in this thesis, several assessment strategies have been used. The accuracy of predictions after training the model has been analysed using metric scores. The Biological significance of found enrichments was analysed using the enrichment score from well-trained models, and then further analysed and visualized in the downstream analyses.

2.3.1 Metric scores

To measure the precision and quality of the training, the R^2 score and RMSE score were used. For ordinal regression, Spearman rank coefficient of correlation was also used, as it deals better with ordinal data.

2.3.1.1 R^2 score

The R^2 score, sometimes referred to as the coefficient of determination, shows how much variance of the true values can be explained by the variance of the independent variable, and thereby gives a probability about how well unseen samples can be predicted. R^2 is defined as:

$$R^2(y, \check{y}) = 1 - \frac{\sum_{i=1}^{n} (yi - \check{y}i)^2}{\sum_{i=1}^{n} (yi - \bar{y})^2}$$

Where yi is the true value and $\check{y}$i the predicted value of the i-th sample from a total of n samples. The term $\bar{y}$ is the average of the true values and is defined as:

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{n} yi$$

(Filho et al., 2011). The R^2 score has an upper bound on 1 but no lower bound. The maximal score can be interpreted as a percentage, where 100% or 1 would mean perfect predictive ability; therefore, the predicted values are exactly the true values. In case of a negative R^2 score the predictions of the model were worse than the mean value, and the mean value is therefore a better predictor than the model itself. Since the R^2score has a positive upper bound, a value close to 1 can always be interpretated as a good result. 100% means that every value was correctly predicted, and the closer to that value we get, the more accurate our predictions (Chicco et al., 2021). As will be mentioned in the following, the RMSE does not have such a clear interpretation by itself.

2.3.1.2 RMSE

The root mean square error (RMSE) gives an idea of the standard deviation of the predictive error, which is the difference between the observed and predicted values. RMSE captures how spread out these values are around the line of best fit. RMSE is obtained by taking the square root of the mean squared error (MSE), another common metric for measuring the accuracy of the model. The RMSE is defined as

$$RMSE = \sqrt{MSE(y, \check{y})}$$

And

$$MSE(y, \check{y}) = \frac{1}{n} \sum_{i=1}^{n} (yi - \check{y}i)^2$$

(Chai & Draxler, 2014). The RMSE score is related to the R2 score, such that with an R2 score of 1 (a perfect correlation) the RMSE will be 0, since there is no deviation from the best fit line of the regression. The RMSE has a lower bound of zero (best result) and no upper bound, and the closer to zero a given result is, the more accurate the predictions of the model, in general. However, in practice, the RMSE score by itself does not give a good indication of the performance of a model, since there is no upper bound and the values could go to positive infinity. A perfect fit is zero, but without knowledge or reference of the maximum RMSE value and the distribution of the true value, judging whether a small RMSE (such as 0.3) is good or bad is not possible, since the scale or range of the RMSE in relation to performance can often not be determined accurately (Chicco et al., 2021). That is why in this thesis, the RMSE score is only secondary and only viewed together with the R^2 score, which as mentioned before, gives a more definitive answer towards performance.

2.3.1.3 Coefficient of correlation

The degree of correlation between two variables, such as x and y, is measured by the coefficient of correlation. It may range from -1 to 1. The value 1 shows that the two variables are moving synchronously and exhibit a positive correlation. They exhibit perfect connection and ascend and descend simultaneously. If the two variables are exactly opposite to each other, the result is -1, and feature a negative correlation. In a perfectly opposing sense, one value goes up as the other goes down. One may argue that any two variables in this universe have a correlation value. If they are not associated, it is still possible to calculate the correlation value, which would be 0. The correlation value is always in the range of -1 to 1, while passing through 0.

There are two widely used types of correlation coefficients, namely Pearson's product moment correlation, for two normally distributed variables, and Spearman's rank coefficient of correlation. The latter is used for skewed and ordinal variables since it deals well with extreme values. The Spearman correlation coefficient is defined as

$$s = 1 - \frac{\left(6 \sum_{i=1}^{n} {d_i}^2\right)}{n\,(n^2 - 1)}$$

where s is the Spearman's rank correlation coefficient, with n samples and d_i is the difference between two sample variable ranks (Mukaka, 2012).

2.3.2 Downstream Analysis

The enrichment score has been developed by Babaei et al. and will be published in a paper in the near future. A draft is included in Appendix I of this thesis (Babaei et al., n.d., Appendix I), thus section 2.4.2.1 is a part of that paper, since this is a unique score calculation without prior reference. The second sub-section will explain the visualization, also derived from the same paper, and explain the theory and intent behind visualization methods.

2.3.2.1 Enrichment Score

We consider two quantities for calculating the enrichment score per cell type, 1) frequency of cell type with high filter response in the spatial neighbourhood of the anchor cells and 2) frequency of cell type with high filter response outside of the spatial neighbourhood of the anchor cells:

We define the score of cell type T to be in the nearest neighbourhood of the anchor cell and selected:

$$S_T^S = \frac{K_T^S}{E_T^S}$$

We define the score of cell type T to be in nearest neighbourhood of the anchor cell:

$$S_T = \frac{K_T}{E_T}$$

where K_T^S is the number of selected cell type T in the nearest neighbourhood of the anchor cell, K_T is the number of cell type T in the nearest neighborhood of the anchor cell, E_T and E_T^S are expected value of cell type T be presented in the nearest neighborhood of the anchor cell and expected value of cell type T be selected in the nearest neighbourhood of the anchor cell, respectively and calculated as follow:

$$E_T^S = K^S \times \frac{N_T^S}{N^S}$$

$$E_T = K \times \frac{N_T}{N}$$

Where the variables K being the number of all cells in the nearest neighbourhood of the anchor cell, N the number of all cells in the image, N_T the number of all cells of cell type T in the image, K^S the number of all selected cells in the nearest neighbourhood of the anchor cell, N^S the number of all selected cells and N_T^S the number of all selected cells of type T.

The enrichment score for cell type T is then defined by:

$$ES_T = \frac{S_T^S}{S_T}$$

The ES values above one indicates enrichment of the selected cells of type CT in the spatial neighbourhood of the anchor cell.

The enrichment score of the global spatial enrichment analysis is calculated by:

$$E_T^S = N^S \times \frac{N_T}{N}$$

$$ES_T = \frac{N_T^S}{E_T^S}$$

The enrichment score for a specific cell type across the patient cohort is reported as the median value of the scores across all patients. The error bars are calculated as the median absolute deviation (MAD).

The enrichment score section is a direct cut-out of the unpublished paper (Babaei et al., n.d., Appendix I).

2.3.2.2 Visualization of the downstream analysis

The downstream analysis will be visualized in three main plots. The first one will show the selected cells by CellCnn and report the Wilcoxon rank sum test p-value, a nonparametric rank test. This p-value depicts whether two populations are evenly distributed based on a rank. The test assigns a rank to values of both populations and the rank mean, or median values of both populations, are compared to one another, to check whether both populations have the same

distribution (Modified Wil, 2012). In the plot described above, the two compared populations are a low and a high functional marker expression.

The main plot will be a logarithmic bubble plot to compare the enrichment between two distinct groups featuring different label expression intensities together with a second plot that will show the calculated enrichment score, featured in section 2.4.2.1, and plot them on the y-axis against available cell types on the x-axis. More details on these procedures will be discussed in the results section. A bubble plot is a type of scatterplot with an additional third dimension, which adds a numerical variable that represents the size of the dots, rather than a third axis. Sometimes even a fourth dimension, colour can be added to represent another variable.

A bubble chart is mainly used to represent and demonstrate correlations with respect to numerical variables. The inclusion of marker size as an additional dimension makes it possible to compare three variables, rather than just two. Three distinct pairwise comparisons can be made: X and Y, Y and Z and X and Z, as well as comparing all three in a single bubble chart, where X, Y and Z represent variables or datasets. To obtain the same number of insights, numerous two-variable scatter plots would be necessary (Brown, 2021). The foundation of the plot are the X and Y axis, these variables are often picked for an existing relationship or to test whether there is a relationship at all between the variables. The gravity of a parameter is represented by the size of the bubble. Bubbles of different colours are used to indicate discrepancies for large collections of data that need to be broken down. These characteristics, which are specific to particular data sets, may be found in any bubble chart (Dougherty & Ilyankou, 2022). In this thesis, bubble charts will be used to compare the logarithmic scale of the enrichment score from two different expression types, high and low, from multiple found cell types found by the downstream, analysis. This can help to conclude whether a certain cell type is truly enriched, compared to other cell types found during the analysis in the vicinity of a functional marker and highlight cells that show a higher-than-expected enrichment.

The logarithmic scale helps in differentiating between actual and just random enrichment, since a difference in this scale is more significant, and illustrates relative, percentage-oriented differences more clearly than the linear scale and deals better with variables that may have drastic magnitude changes.

The logarithmic bubble plot presented in this thesis and the visualization used in the downstream analysis have been derived from the unpublished paper (Babaei et al., n.d.), found in Appendix I of this thesis.

3. Results

In this chapter, some prior CellCnn analysis was performed to show that the framework works and assessed the general predictive power of CellCnn. Then, to select the best correlation, a few reasonable cell types have been chosen and compared to different functional markers. Once a set of well-performing cell types has been identified, different ways of computing the functional marker labels for those cell types was tested, to further improve the results. Finally, before performing the downstream analysis where the enrichment of the best results was computed, the current approach was optimised. To achieve that, the k-nearest neighbours have been examined further, by constructing a trend plot from k=10 to k=50 in steps of 10.

Due to the time constraints of this project, only the 3 best performing and optimized results will be further analysed with the downstream analysis.

Finally, some further approaches will be mentioned, and their results depicted here.

This section will only focus on showcasing all the results obtained in this thesis and the settings used to obtain them. All major interpretations will be featured in the discussion in Chapter 4, where the results will be critically analysed and compared to relevant literature.

3.1 Testing CellCnn

3.1.1 Linear toy model

To test CellCnn in a more generalized setting, a linear toy model was constructed and compared to a standard neural network from TensorFlow. Both used linear regression as the only activation function.

The data has been created using the library scikit-learn with its *make regression* function, which creates a random regression problem. To get closer to the structure of the CellCnn problem, the data included 6000 points for 55 features to create an input set of dimensions 6000x55, a multiparametric regression problem. The created features correspond to the functional markers in CellCnn. TensorFlow linear neural network was compared against CellCnn with linear activation. Figure 5 below shows the comparison for the correlation of the TensorFlow linear model with the CellCnn linear model.

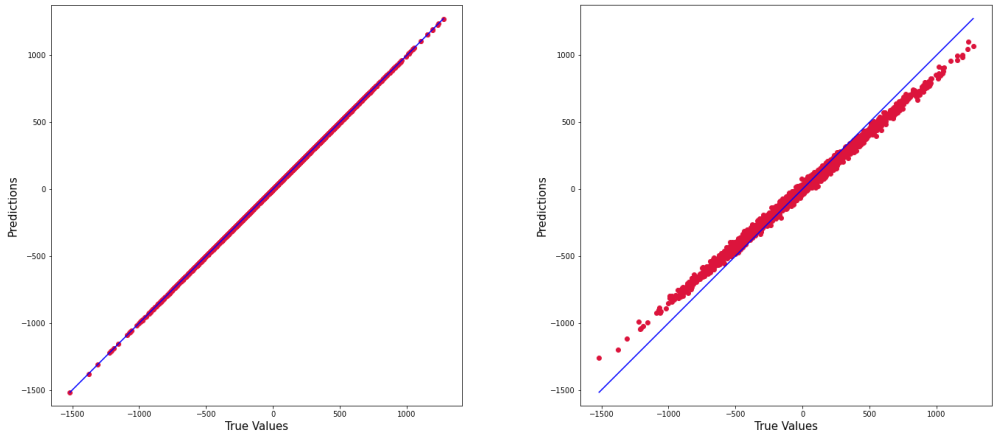


Figure 5: Shows the correlation plot of the TensorFlow linear model on the left and CellCnn on the right. The true values were plotted against the predicted values. The R2 score of TensorFlow was 99.89% and the R2 score from CellCnn was 97.5%. CellCnn is able to detect simple linear patterns.

3.1.2 Frequency analysis

In order to test CellCnn's predictive power further, a more complex and thematically more appropriate problem has been analysed. Like in the main analysis, the frequency of a target cell has been computed around an anchor cell, such as all CD4+ T cells. This frequency is then used as a label to train the model, instead of the functional markers. The aim of the frequency analysis was to fit the probability distribution of the occurrence of a target cell type in the proximity of an anchor cell. The first setup shows the frequency of CD4+T cells in the vicinity of granulocytes considering both groups of patients, taking the respective regions into account. This represents one of the key findings in the paper by Schürch et al. (2020), who detected an

enrichment of PD-1+ CD4+ T cells around the neighbourhood of granulocytes. The result was compared for both groups and each group individually. The k-value has been set to 50, since some regions contain only 80+ cells entirely, and the number of available cells varies greatly; the number was kept lower to allow for better distribution of values. Table 2 below shows the metric score of the first results.

Table 2 shows the frequency analysis for the anchor cell set for all types around the vicinity of granulocytes with the nearest neighbour k set to k=50.

Group	R^2 (%)	RMSE
1	25.9	4.1
2	28.5	2.3
Both	52.4	2.4

Another case, not directly related to the paper from Schürch et al. (2020), was also analysed by looking at the frequency for CD68 + macrophages in the vicinity of granulocytes, again for both groups, then each group individually. The results were much worse than the previous example. Table 3 shows the corresponding metric scores.

Table 3 shows the frequency analysis for CD68 + macrophages in the vicinity of granulocytes with k=50 nearest neighbours.

Group	R^2 (%)	RMSE
1	6.4	1.9
2	2.9	2.3
Both	7.1	1.9

Since the results were generally not as good as expected, based on the findings of the paper, for example, the frequency analysis was not further deployed and will be further investigated in the discussion.

3.2 Anchor-based functional Spatial enrichment analysis

For the functional spatial enrichment analysis, multi-cell input was generated (observations of cellular populations) using nearest neighbour cells for each patient from a chosen anchor cell, with a set of marker expressions as features, that aim to capture functional activity. Since the neighbourhoods were constructed around an anchor cell, computing the closest neighbours k, the aim was to identify the enrichment in spatial proximity of a specific anchor cell type of cells which express a chosen functional marker.

During training, the data was separated into their respective survival groups, and each patient was analysed individually prior to merging. A cell type was chosen as an anchor cell, then the neighbourhood around each anchor cell in that patient was computed, while taking the respective TMA (tissue microarray) region into account. Thus, only cells in the same region were considered for the neighbourhood. A functional marker was decided on as a label for each training attempt. Cell identifiers and the marker intensity of the labels around that neighbourhood were documented for further analysis and linked to all k-nearest neighbours of the chosen anchor cell, using all available markers (minus the label) as a dependent variable, and the label intensity as an independent variable for training. This process was repeated for all anchor cells in a patient, then for each patient. The gathering of these values for all patients around the chosen anchor cell, given a functional marker, will be referred to as data in the following. For training, the data consisting of the neighbourhoods was split into training and test data, where the test data was entirely excluded from training. Training data was further split into 80% training data and 20% validation data, used to estimate the predictive power of the model during training. The concept of CellCnn is depicted in Figure 6 below.

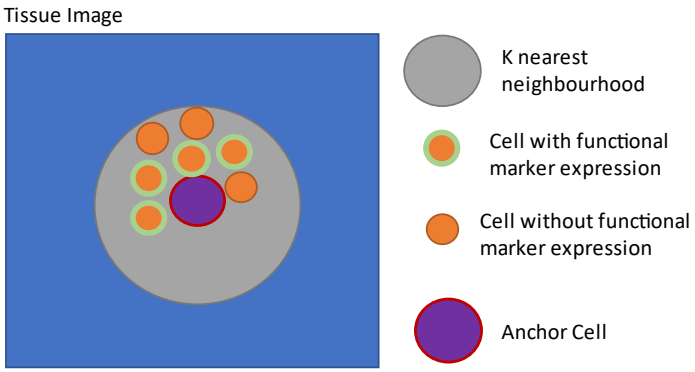


Figure 6 Illustrates a simple concept of a functional and spatial enrichment analysis in the vicinity of an anchor cell. The k-nearest-neighbours are computed around the anchor cell considering only cells with the chosen functional marker expression.

3.2.1 Comparing different cell types with different functional markers

The goal of this sub-chapter is to identify a good starting point for further analysis, by testing several cell types against different functional markers. The chosen markers are explained in the method section in Table 1, and consist of EGFR, ICOS, KI67, Beta-Catenin and BCL-2. Each of the results in this chapter were produced with some fixed parameters: the activation function of CellCnn was ReLU activation function (section 2.2.4), for all but the final layer which used a linear activation function. 200 iterations were fixed during training, binning was set to 5 bins and k was fixed to k=10 for the k nearest neighbour computation. The main cell types focused on in this thesis are B cells, T regulatory cells, CD4+ CD45RO T cells, CD8+ T cells and granulocytes. The cell types were chosen from the original datasets based on their biological significance during cancer progression and from a computational point because of the number of available cells for that cell type. For example, CD4+ T cells only has 2303 cells, while CD4+CD45RO+ T cells have 16661 representative cells. In order to gain good results only cells with sufficiently high enough amounts were chosen, while the variety was further limited to those 5 cells because of time constraints, since optimization becomes much more time consuming per added cell type.

Figure 7 below, shows the result of Ki67 marker with the R^2 score (R2) and RMSE score.

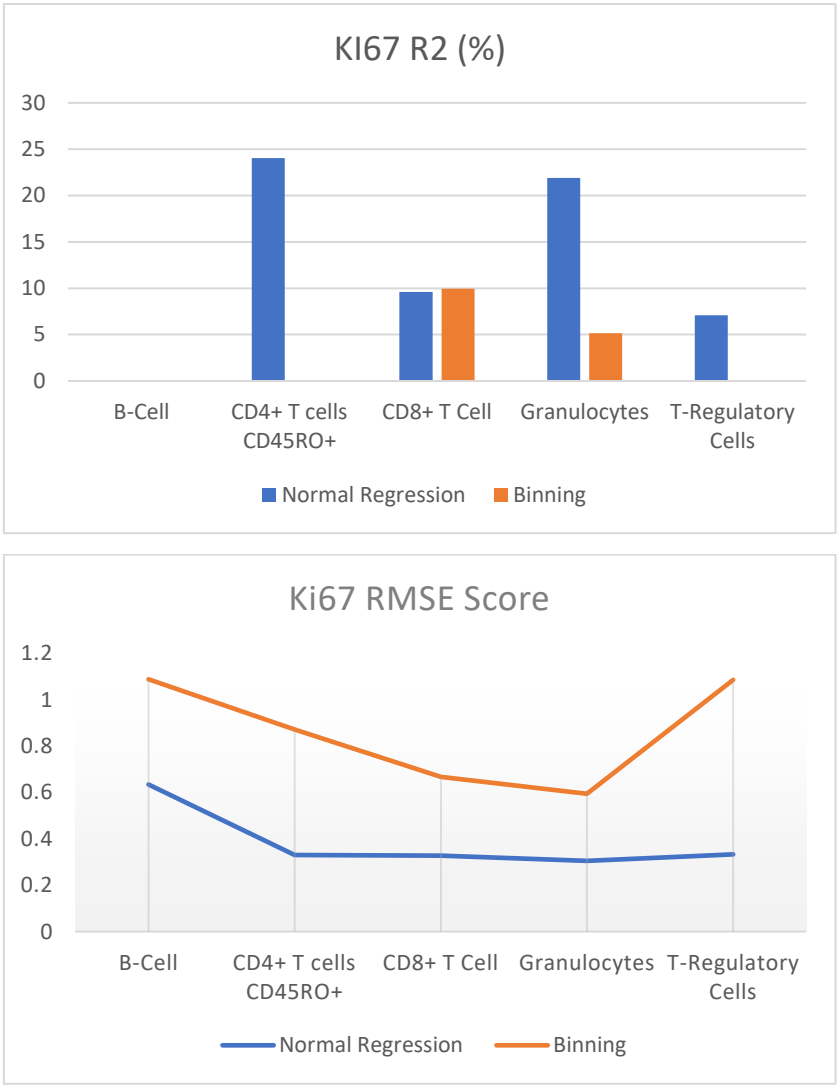


Figure 7: Ki67 proliferation marker with R2 and RMSE score of different cell types.

B cells had a negative score for both normal and binned regression, while CD4-Cd45RO T cells as well as regulatory T cells got negative binning results, those were excluded from the plot since they give no meaningful result and failed during training. Overall, Ki67 gives very poor results, especially compared to the following results below. The same goes for the RMSE score, typically close to zero, shows many values around 1 or even higher. Thus, Ki67 will be excluded as a marker for further analysis.

The Figure 8 below, shows the metric scores for EGFR functional marker with several cell types.

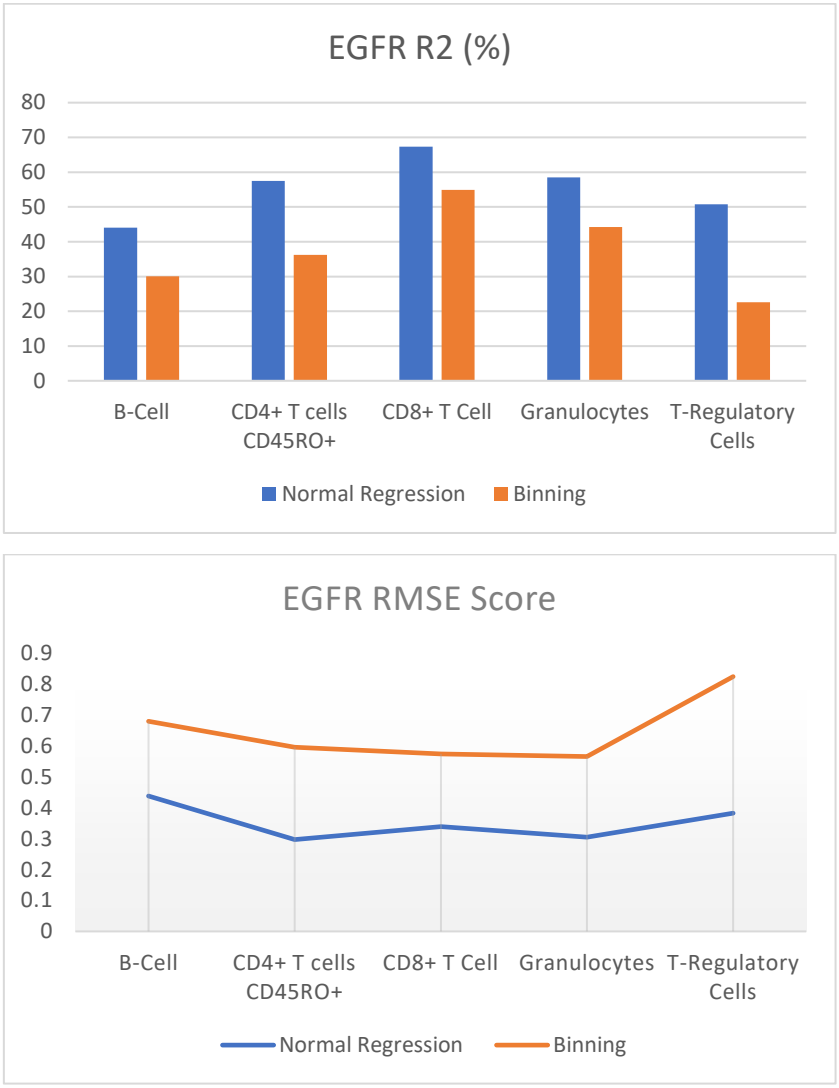


Figure 8: Epidermal growth factor receptor (EGFR) marker with R2 and RMSE score.

The marker for EGFR shows very stable results for RMSE without binning, whereas binning has many results closer to 1. Thus, the normal regression appears to be more stable than the binned approach. The R2 score is also always lower for the tested cell types with binning compared to normal regression. CD8+ T cells seem to give the best result so far, for both binning and regular regression.

Figure 9 illustrates the metric scores for the BCL-2 marker for several cell types.

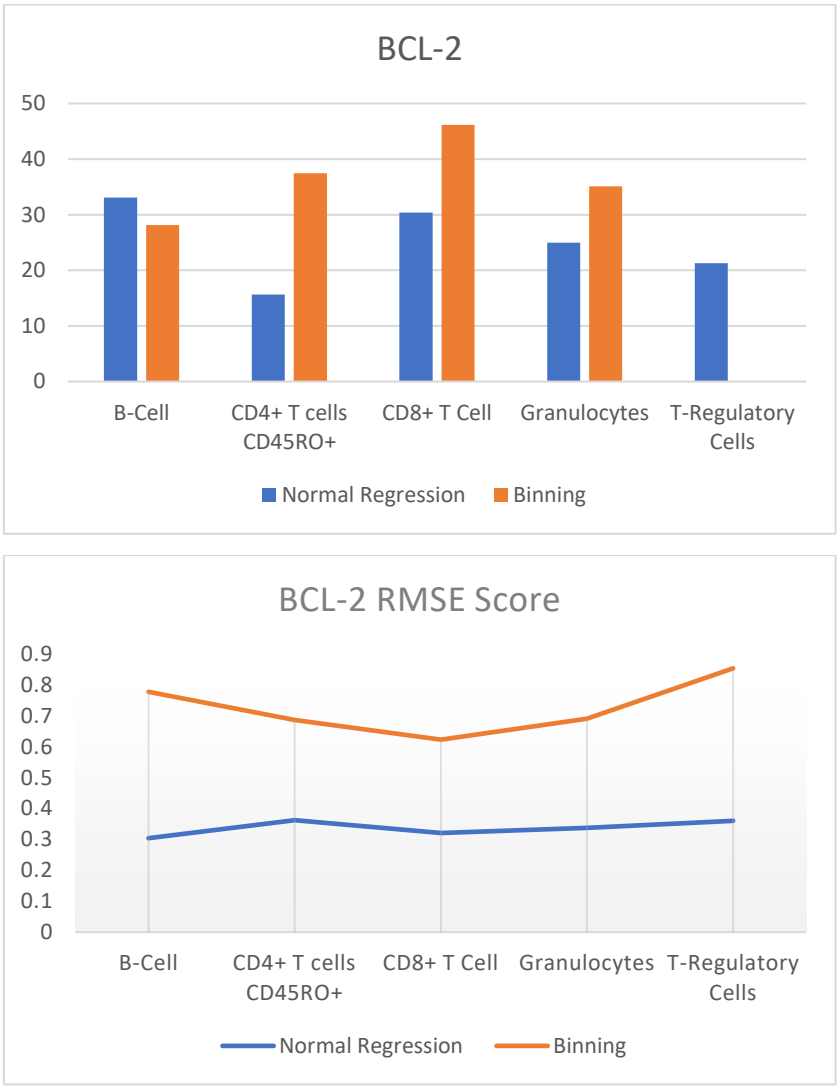


Figure 9: BCL-2 functional marker with different cell types, comparing R2 and RMSE score.

BCL-2 again shows high RMSE scores for binning and acceptable scores for normal regression around 0.3 ± .05. CD8+ T cells scored with the highest R2 score, but the results are generally lower than for the EGFR marker.

The metric scores for R^2 and RMSE for Beta-Catenin are depicted in Figure 10 below.

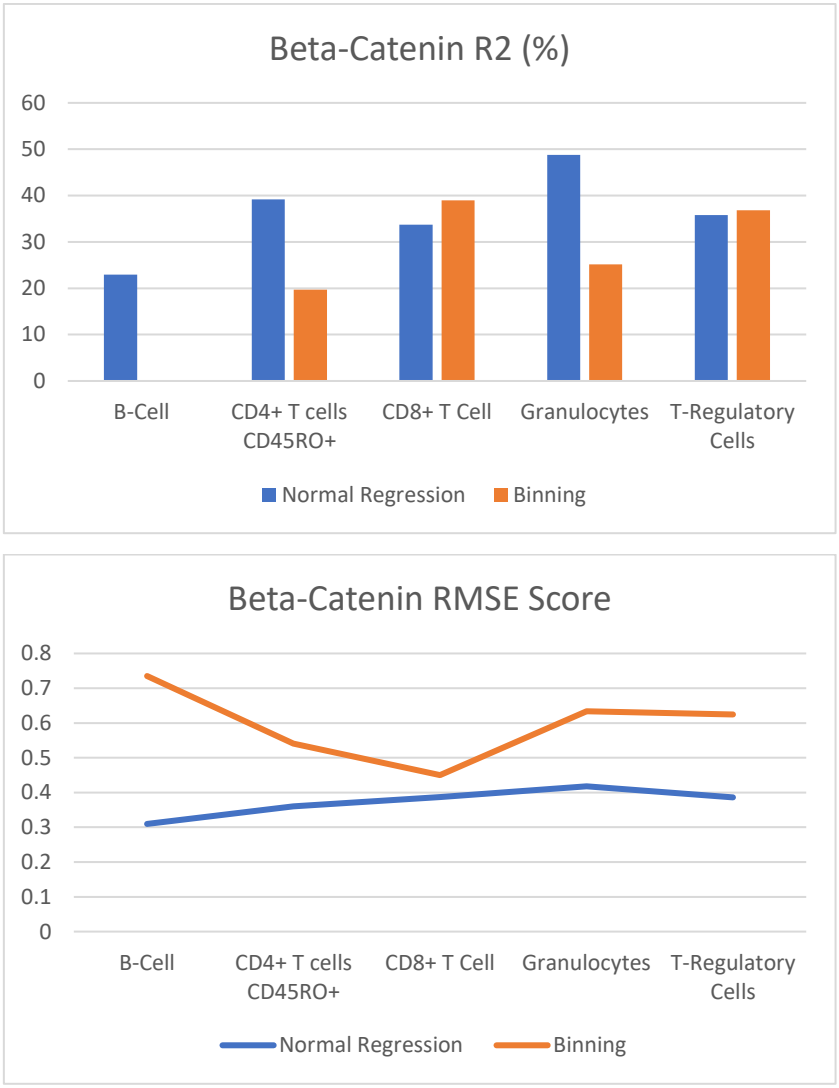


Figure 10: The next chosen marker of, the beta-catenin marker different cell types computing the R2 and RMSE scores.

Beta-catenin performed best for granulocytes. Although the RMSE is highest for normal regression for granulocytes, the score is still close to zero and has an acceptable R2 score. Binning shows very high RMSE scores and lower R2 scores, for B cells even showed negative R2 scores. CD8+ T cells show the lowest RMSE score for binned values but does not perform well with R2 scores below 40% for either performance.

The ICOS functional marker and its metric scores for comparing several cell types is shown in Figure 11.

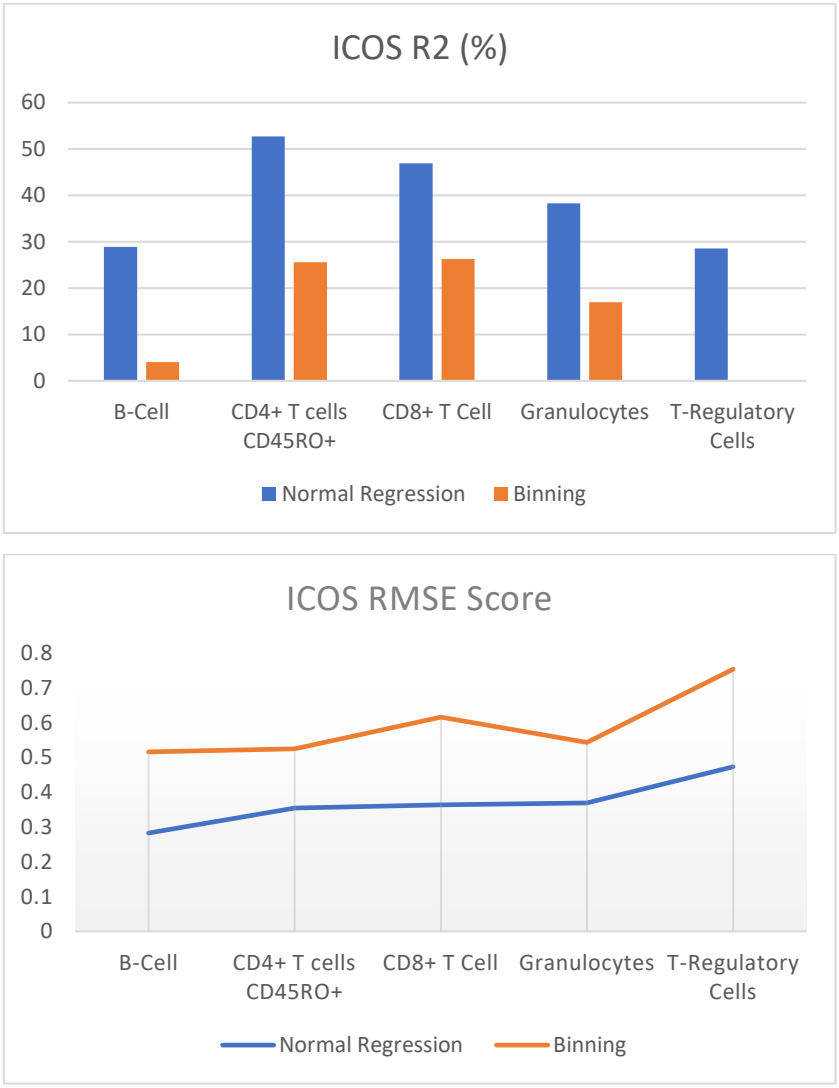


Figure 11 shows R2 and RMSE plots for different cell types using ICOS marker. For ICOS the CD4-CD45RO T cell type performed the best.

Once again, binning shows lower R2 and higher RMSE scores. Since binning consistently performed worse than regular regression, it will be excluded for future analysis, and for the main results of this thesis, regular regression with CellCnn will be used as a fixed parameter. Often, for different functional markers, different cell types seem to perform better for normal regression. For binned data values CD8 + T cells consistently performs the best.

Table 4 and 5 below show the best performing cell types for normal regression and binning respectively.

Table 4: Best Normal Regression Performance with the training label, best-performing cell type and the R2 and RMSE Score

Label	Best Cell Type	R2 (%)	RMSE
KI67	CD4+ T cells CD45RO+	24.04	0.33
ICOS	CD4+ T cells CD45RO+	52.71	0.35
EGFR	CD8+ T Cell	67.34	0.34
Beta-Catenin	Granulocytes	48.76	0.42
BCL-2	B-Cell	33.11	0.31

Table 5: Binning best performance with the training label, best performing cell type and the R2 and RMSE Score

Label	Best Cell Type	R2 (%)	RMSE
KI67	CD8+ T Cell	9.93	0.66
ICOS	CD8+ T Cell	26.26	0.61
EGFR	CD8+ T Cell	54.93	0.57
Beta-Catenin	CD8+ T Cell	38.99	0.41
BCL-2	CD8+ T Cell	46.15	0.62

Table 4 and 5 show the best results for normal and binned regression respectively. The worst results come from Ki67, the functional marker. All not included bars in the illustrations above had negative R2 score, and therefore performed worse as predictors than the mean value. This could suggest that Ki67 may not provide a pattern or correlation to learn from and, as stated, will be excluded from future analysis. Some binning results, such as EGFR with CD8+ T cells had acceptable R2 scores, however as mentioned above, due to the higher RMSE scores and generally lower R2 scores, binning will no longer be used for the main task of this thesis, to allow for more efficient general optimization. Since the metric scores of BCL-2 were not as high as for other functional markers, it will be excluded for the following optimisations, due to time-constraints. Thus, EGFR, Beta-Catenin and ICOS will be analysed in the following, using only the regular regression option with CellCnn.

3.2.2 Optimizing the best-performing results

To increase the best results of the previous step, multiple different strategies have been implemented. For this step, CD4+-Cd45RO+ T cells with ICOS, granulocytes and CD8+ T cells with the EGFR and granulocytes with beta-catenin functional marker have been selected from the best results shown in the previous section 3.2.1.

3.2.2.1 Different label computations and different number of markers

The different ways to compute the labels using minimum, maximum, average, median and percentile of maximum were tried out to test for better ways of computing the labels for training, to further improve the results. The best results from before were tried with different computation modes, while keeping the best label and the k nearest neighbours (k = 10) fixed: Average (avg), maximum (max), minimum (min), median (med), 20% minimum percentile (nm20) and 80% maximum percentile (nm80). For the percentile computation, the nearest neighbours were ordered from lowest to highest value, and the cell at the index corresponding to the percentile was chosen. For example, for an 80% maximum percentile with k=10, the index would be the eighth entry in the list. Since granulocytes with EGFR is still higher than most of the best results up to this point, granulocytes with EGFR were also tested with the different label computing methods and additionally tested with all 56 markers and the regular 48 markers as a comparison.

Figure 12 shows the comparison for different label computations and marker sizes for EGFR with the anchor as granulocytes.

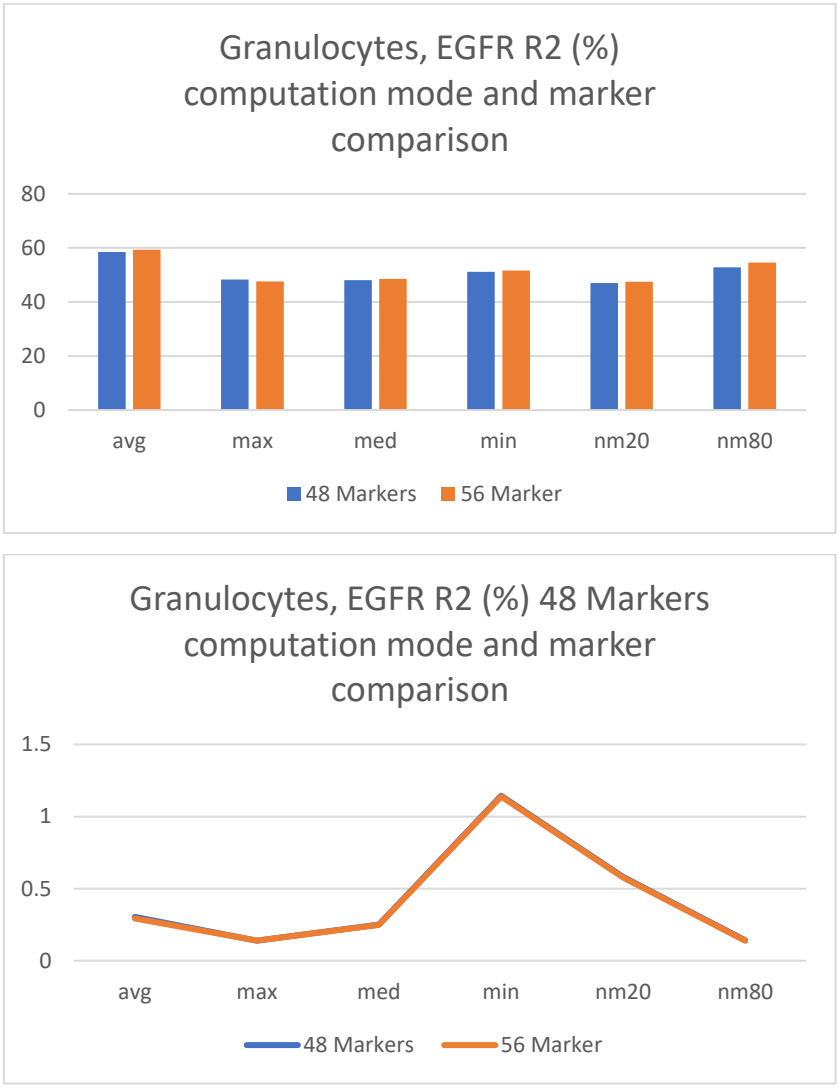


Figure 12: Comparison between different computation modes of the EGFR label with 56 and 48 markers of granulocytes.

The best result used all markers with the average computation of the label with an and R2 score of 59.3 % As can be seen in Figure 7, the results are very similar, almost identical for the different marker sizes. However, 56 markers are slightly higher in their R2 score.

The different label computations and number of markers for EGFR with CD8+ T cells as anchor cell, are compared in Figure 13.

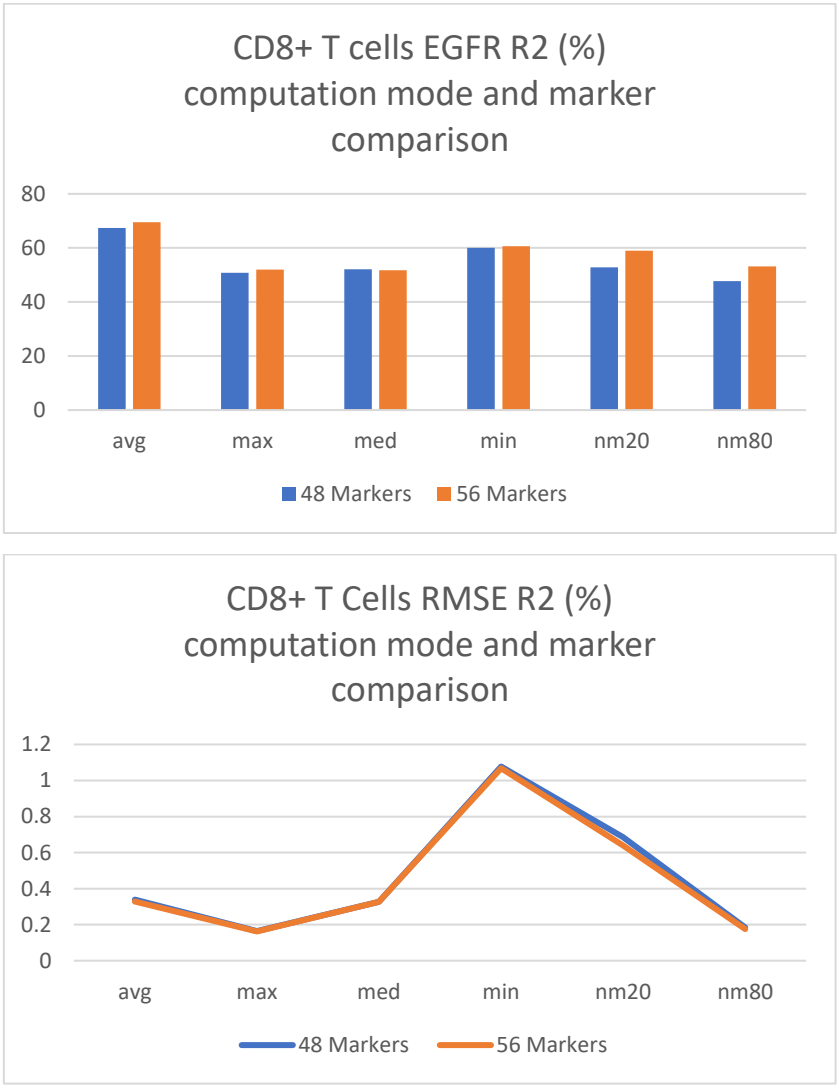


Figure 13 illustrates the comparison between the different modes of calculation of the EGFR label with 56 and 48 markers of CD8 + T cells.

The best result coming from the average label computation with 69% R2 score used all 56 markers. The results are often very similar and produce comparable scores, but again including all markers slightly improves the R2 score.

Figure 14 shows the CD4+ CD45RO+ T cells as anchor cell, comparing different ICOS computations, as well as different number of markers.

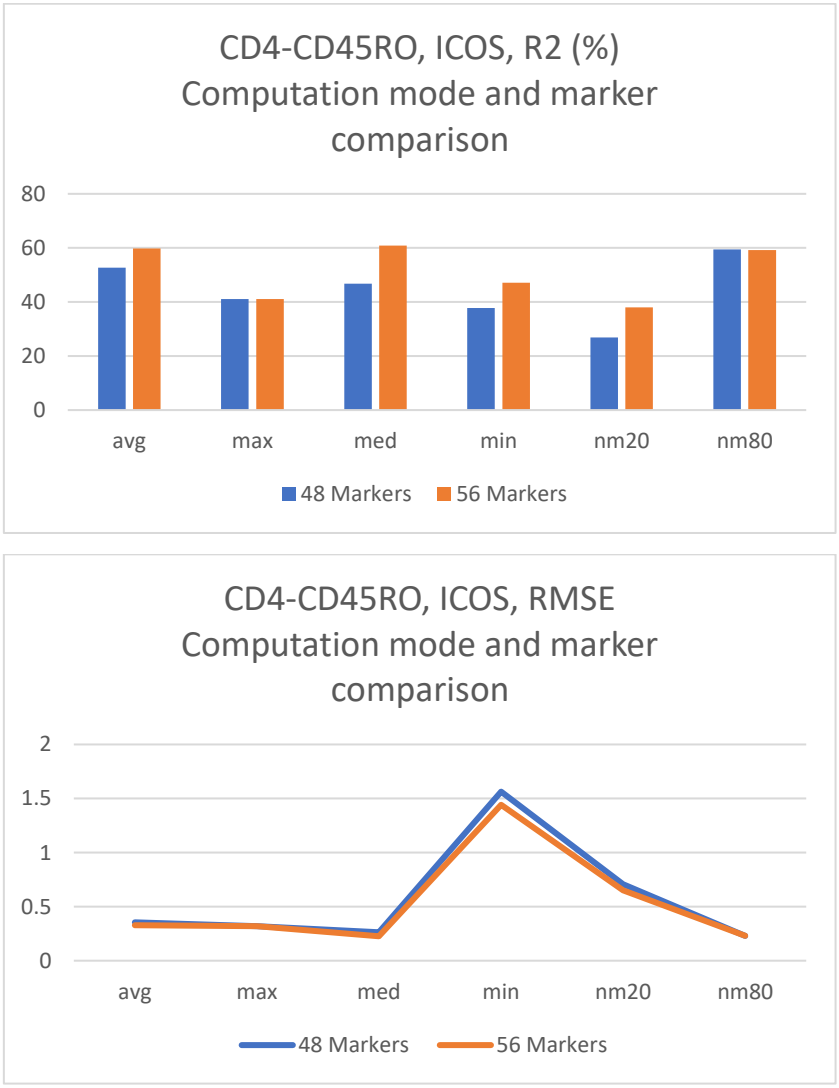


Figure 14 shows the comparison of 48 and 56 markers with different computation modes, for CD4-CD45RO T cells.

The highest result was achieved using all markers with an average calculation with an R2 score 59.7%. The RMSE score between the two marker amounts is essentially the same, while the minimum shows the highest RMSE score.

Figure 15 below, compares different marker sizes and computation modes for Beta-Catenin, using granulocytes as an anchor cell type.

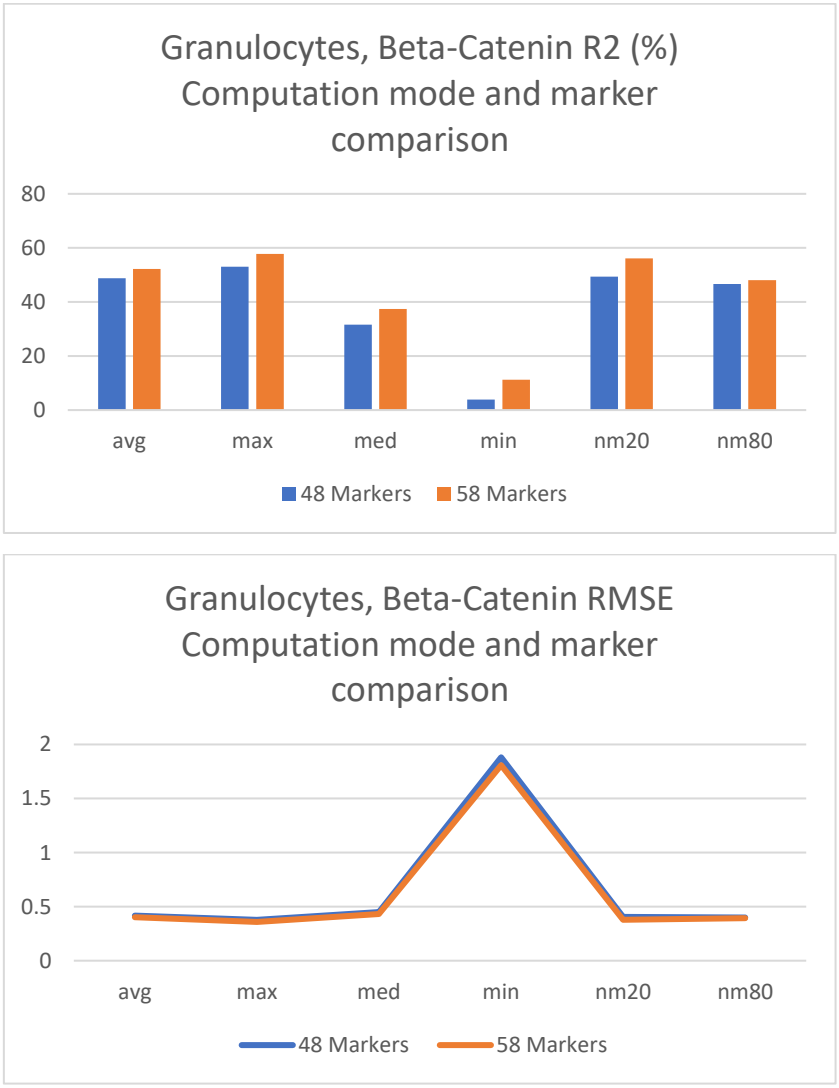


Figure 15: Granulocytes with beta-catenin. Comparison of different label modes and maker size.

The highest R2 score came from the maximum label computation using all markers with 57.7%. Since this result is lower than granulocytes with EGFR, and because both results had the same cell type as the best result, this result will not be further improved or considered for the Downstream analysis.

3.2.1.2 Trend plots k=10-50

The k-nearest-neighbour value k was computed in steps of 10 from k=10-50. For CD8+ T cells and granulocytes with EGRF, as well as CD4+CD45RO+ T cells with the ICOS marker, the trend plots are shown together with their corresponding correlation plots in Figure 16 a-f below.

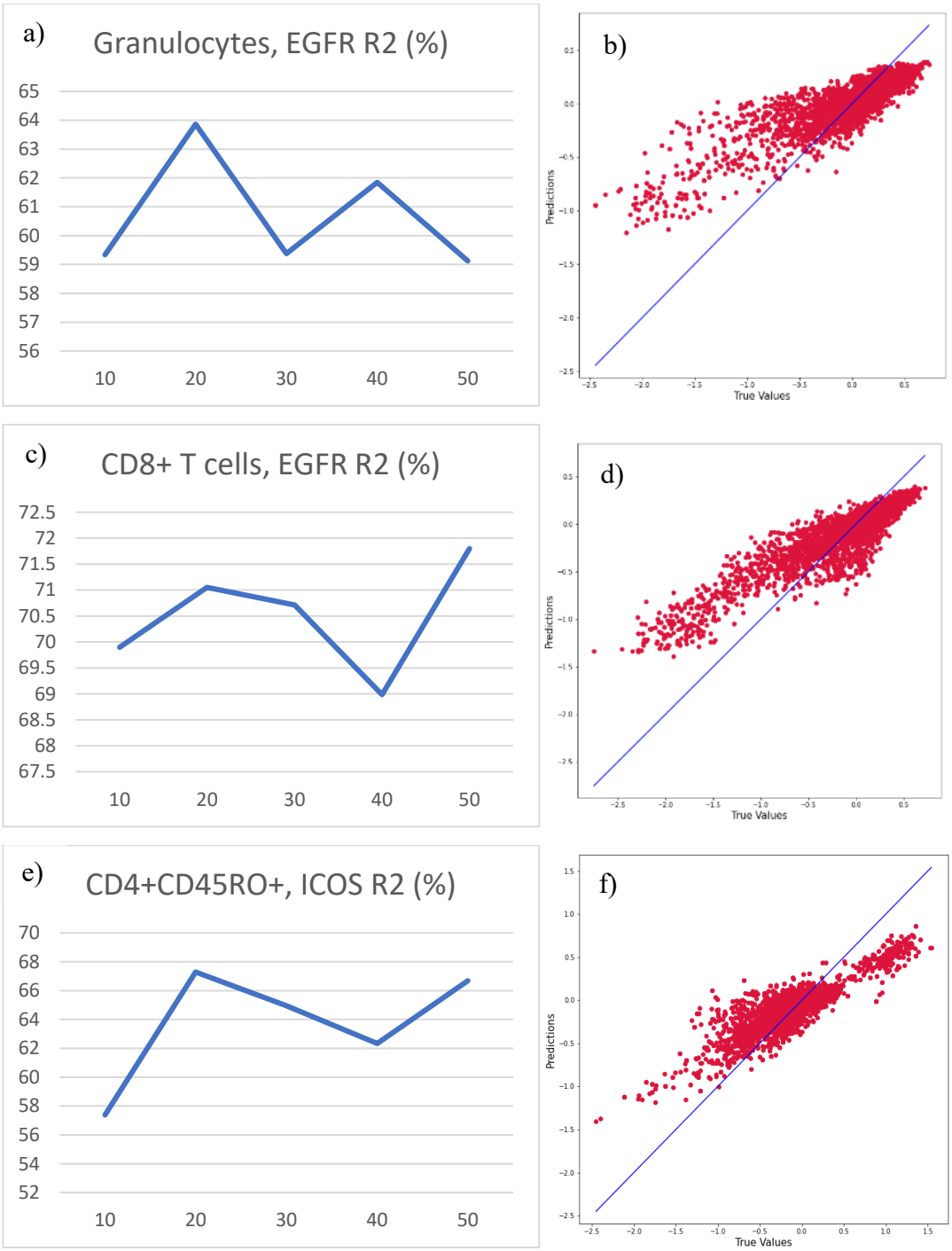


Figure 16 shows the trend plot for increasing k values for granulocytes, CD8+ T cells, and CD4+CD45RO+ T cells in Figures 11a, 11c and 11e, respectively, while Figures 11b, 11d, and 11f show their corresponding correlation plots in the same order, with the true values plotted against the predicted values. The left side of each trend plot shows their corresponding correlation plot.

For the granulocyte neighbourhood in the proximity of the EGFR marker, the result was improved by k = 20 to 63.8%, as shown in Figure 16a. CD8+ T cells had the highest result with k=20 nearest neighbours with 71%, although k=50 is slightly higher (71.8 %), the computational cost given the possible error range (random factor during training) means k=20 is computationally the more efficient solution. Figure 16c shows the increasing k-values for CD8+ T cells. CD4+CD45RO+ T cells with the functional marker ICOS depict in their trend plot, in Figure 16e, that the highest score is 67% with K=20, increased from 57%.

For all trend plots the RMSE score was at 0.3+/-0.5. It remained stable throughout the trend plots for the same label computation mode. The RMSE score was very consistently worse for the minimum computation mode compared to the others.

3.2.3 Downstream Analysis

In the downstream analysis, the results of training CellCnn, and the filter responses are analysed and linked to all available cells in order to test for the corresponding enrichments. Downstream analysis will only be performed on the best results from 3.2.1.2, the 3 best settings. For the downstream analysis, the neighbourhood of an anchor cell in relation to the functional marker label will be investigated. Each of the 258386 available cells has a corresponding filter response (weights) to them, the list was filtered based on two factors: whether the cell is present in the neighbourhood in question and whether it was above an arbitrarily defined threshold, which excludes cells with a lower response than the threshold.

3.2.3.1 Filter selection

CellCnn uses two filters that potentially learn different patterns from the dataset. These two filters are capable of learning a different pattern or correlation about the given data. More or less filters could be chosen during training, the two filters are an arbitrary number. This thesis used two filters to allow for some variety during training, while still keeping the complexity lower. The final parameter for the downstream analysis was a separation of the selected cells into a high functional marker response in the vicinity of the anchor cell and a low expression, which can be chosen predominantly by a filter based on the learned pattern. This will be the main difference for the two filters, whether they picked cells with the high or low expression preferably. This will be shown in the following plots.

3.2.3.1.1 Granulocytes with the EGFR marker

Figure 17 shows the filter selection for granulocytes with EGFR as label, for both filters.

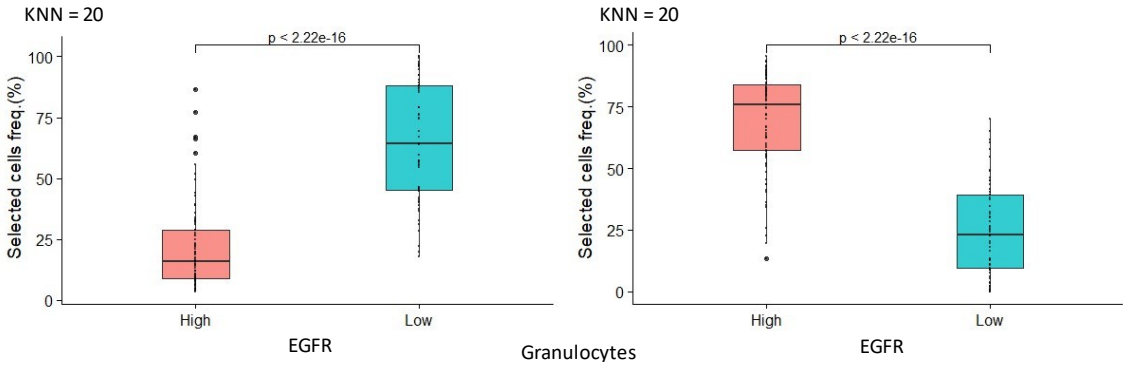


Figure 17 shows the filter selection for the EGFR label in the vicinity of granulocytes. The first filter predominantly selected low expression cells while the second filter chose high-EGFR label expression cells more frequently.

3.2.3.1.2 CD8+ T cells with EGFR marker and k=20

Figure 18 below shows CD8+ T cells with EGFR as label for the two filters.

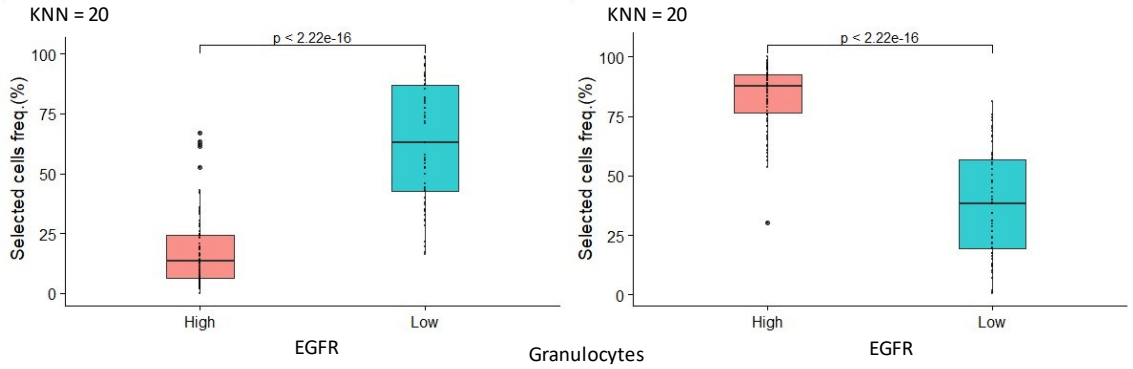


Figure 18 shows the filter selection for the EGFR marker in the vicinity of CD8+ T cells. Similar to the first case in 3.4.1.1 the first filter selects low-label expressions more frequently than the second filter.

3.2.3.1.3 CD45+CD45RO+ with the ICOS marker and k=20

The selection for both filters for CD4+CD45RO+ T cells together with the ICOS functional marker is shown below in figure 19.

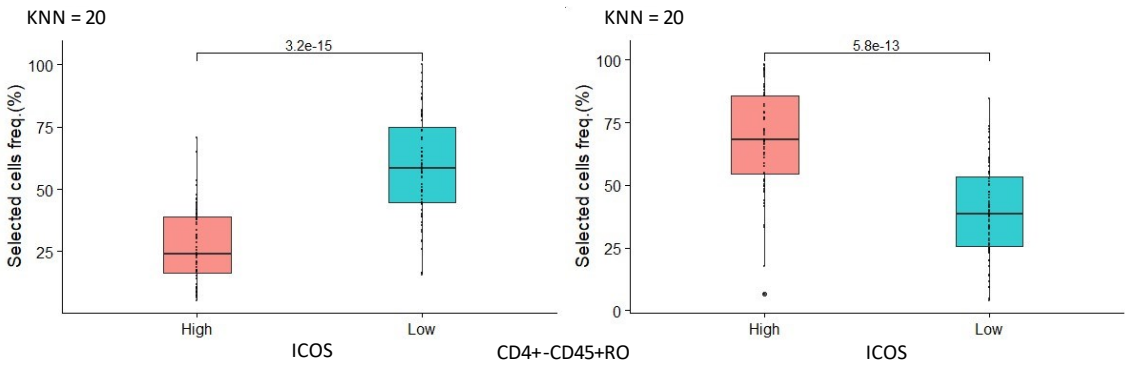


Figure 19 shows the filter selection for the ICOS marker in the vicinity of Cd4-Cd45RO cells. The left illustration shows the first filter, which like the previous cases selected cells with a low expression, while the second filter on the right, picked cells with a high ICOS expression more frequently.

3.2.3.2 Logarithmic enrichment bubble plots

The bubble plot, introduced in chapter 2.3.2.2, depicts the logarithmic enrichment score for the low expression on the x-axis, and the high label expression on the y-axis. A trend line is visualized to aid in the interpretation of the results. Only cells in the positive quarters of the plot and not on the trendline would have a higher-than-expected enrichment compared to the other group. Negative scores on either axis and scores on the trendline can be ignored as they either do not show an enrichment or significant difference, respectively. Positive outliers offer crucial information for this thesis because they show enrichment that differed more than "expected" from the enrichment of other cell types. Each bubble also features an error bar, to further aid in analysis and interpretation.

Next, the enrichment scores were calculated for both filter selections as described in the method section 2.3.2.1. As mentioned above, every cell in the dataset has a corresponding filter response, and only filter weight responses above the threshold were selected. Furthermore, the cells were further filtered to be in the same patient, in the same survival group, in the same TMA region and the same neighbourhood around an anchor cell. The enrichment shown below shows the bubble plot and individually plotted low and high marker expressions plotted against the selected cell types for each filter, to aid in interpretation and give a general overview.

An important aspect in interpretation will be the selected label expressions, high or low, shown in section 3.2.3.1. If a filter only selected cells, around a neighbourhood of a given high

expression marker, then enrichments in the bubble plot with only a low expression, positive enrichment score may not be reliable since they may be underrepresented.

3.2.3.3.1 Granulocytes with EGFR

The following results are in the vicinity of granulocytes considering an EGFR functional marker with high or low expression. Figure 20 a-f illustrates the bubble plots for both filters with their corresponding bar plots, showcasing the selected cell types with their enrichment score.

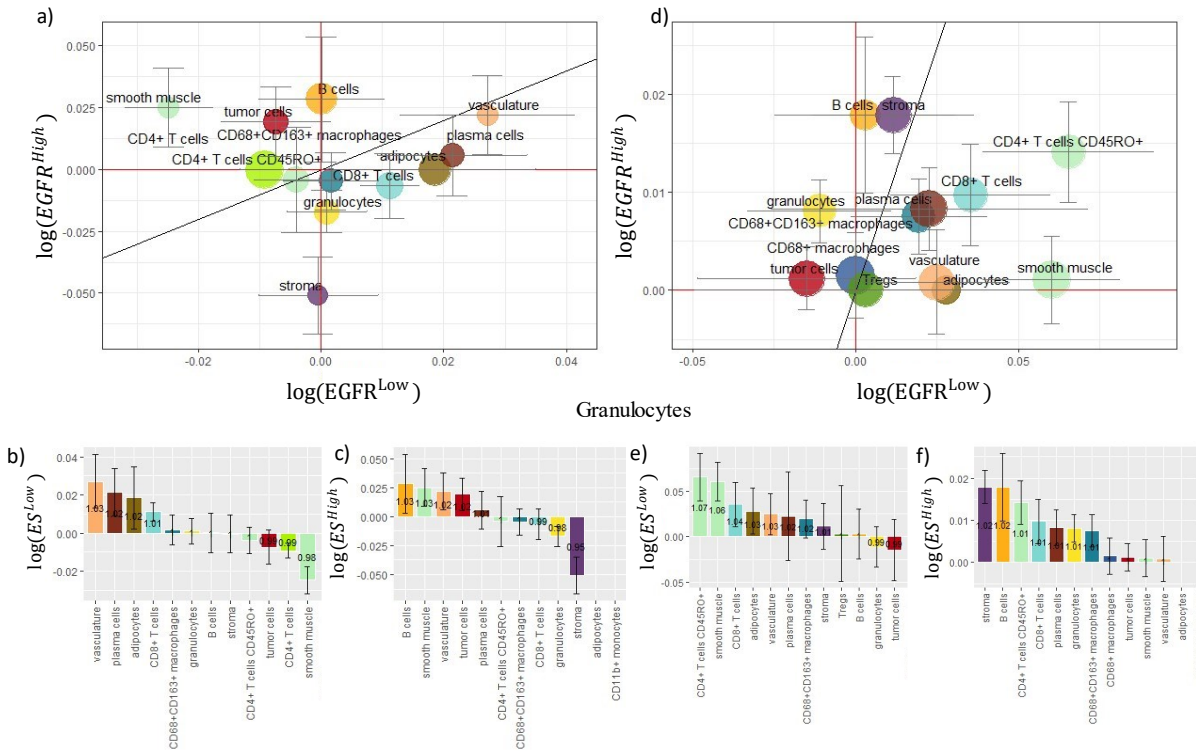


*Figure 20 shows the bubble plot for the first filter **a)** and the second one **b)** on the right top, with **b)** and **c)** showing the logarithmic low and high label expression plotted against the selected cell types respectively for the first filter, and the same for the second at **e)** and **f)**.*

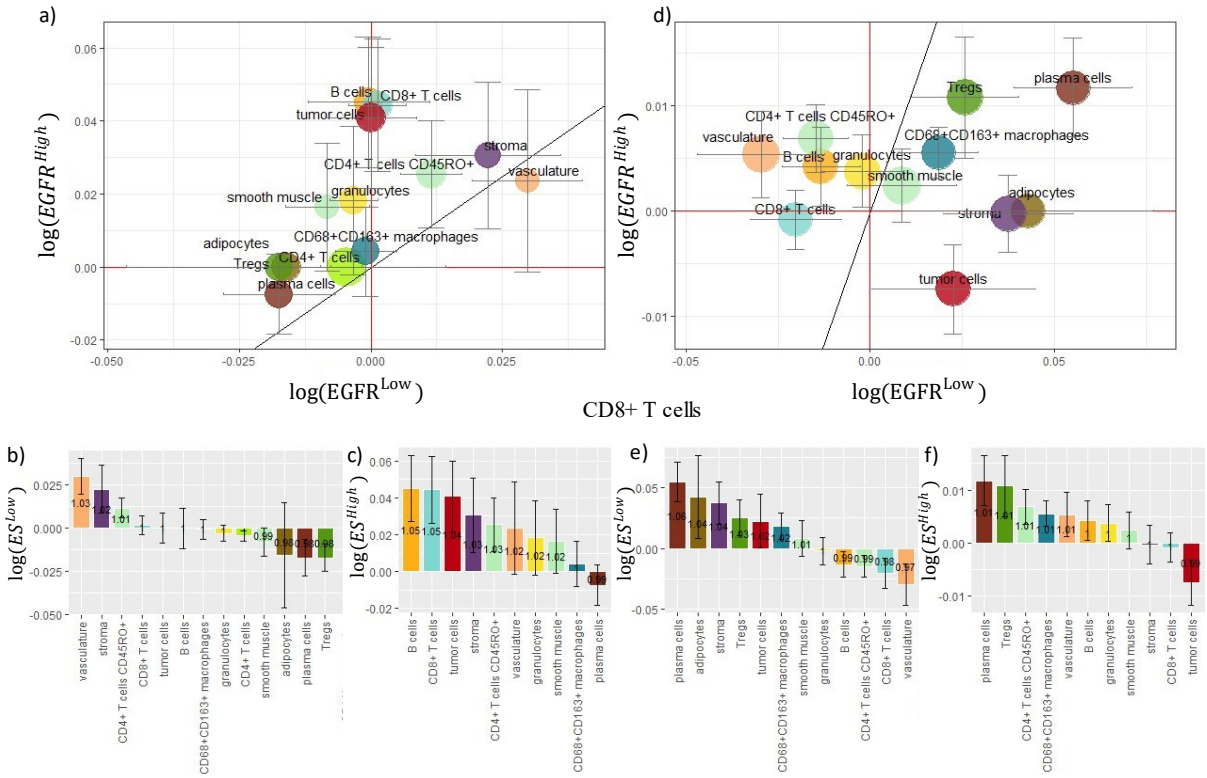
The first filter, which selected low EGFR expression cells, shows a high enrichment of B cells, tumour cells and smooth muscle cells in the presence of high EGFR expression but not the low expression, which can also be seen in Figure 20c showing higher positive logarithmic enrichment scores that are very small in the low expression, seen in Figure 20b. Since those are in the high expression group for a filter that selected mostly low expressions, these results are not very reliable and will not be regarded for the discussion. Plasma cells and vasculature cells show an enrichment in both low and high expression as well as the corresponding values

in the logarithmic plots in 20b and 20c, while the low expression by itself only shows an enrichment in adipocytes and CD8+ T cells, where adipocytes do not appear in Figure 20c, the high expression plot, at all. Vasculature cells lie on the trend line, thus suggesting a similar enrichment in both groups, from which no further information can be extracted. There appear to be no selected adipocytes for high expression of EGFR in the vicinity of granulocytes, and CD8+ T cells has a lower value whose error bar goes into the negative quadrant on the plot in Figure 20c, while the corresponding bubbles in Figure 20a are negative for high but not the low EGFR expression. Therefore, an enrichment in CD8+ T cells, adipocytes, and plasma cells in the vicinity of low expression of EGFR around granulocytes can be concluded.

The right part of the Figure, 20d depicts the second filter, which selected high EGFR expressing cells more frequently. Figure 20d shows an enrichment of CD4-Cd45RO T cells, CD8 + T cells and plasma cells present in the positive quadrant of the plot and below the trendline, thus suggesting an enrichment in both expression groups. This is further confirmed in Figure 20e and 20f with both showing those cell types in their positive region and with relatively large peaks. However, the bubble plot (20d) clearly shows that those cells are enriched more in the low EGFR expression, and since this is a high expression filter, these results cannot be further considered. B cells and stroma cells appear to be enriched more around the high label expression, where B cells have a very low peak in Figure 20e for the low expression, and stroma cells also show a small peak with an error bar for both that reach into the negative values. This can be further verified by stroma cells being close to and their error bar reaching past the y-axis and the trendline in Figure 20d. Smooth muscle cells have a larger positive peak in Figure 20e, for the low label expression and a value close to zero for the high expression, thus there seems to be an enrichment of those cells for the low label expression only. Vasculature and adipocytes also show a smaller enrichment for the low expression only and thus, those cell types will not be considered as well. The other cell types can be seen in the bubble plot of (20d), close to the trendline, thus not showing a significant enrichment in comparison. The main finding of filter 2 is an enrichment of B cells in the high expression proximity of granulocytes.

3.2.3.3.2 CD8+ T cells with EGFR

This section shows bubble plots for both filters trained with cellular neighbourhoods in the vicinity of CD8+T cells around the EGFR functional marker, shown in Figure 21 a-f.



Figures 21a and 21d show the bubble plots, depicting the logarithmic low marker expression of the enrichment score on the x-axis and the high marker expression on the y-axis, for the first and second filters, respectively. Figure 21b and 21c show the individual logarithmic enrichment score of low and high expression respectively, for the first filter, while Figure 21e and 21f show the equivalent plots for the second filter.

Figure 21a, the first filter detected an enrichment in cell types such as granulocytes, Smooth muscle cells, and B cells have positive values for both high and low expression, but the scale shows a higher enrichment in the high expression filter. As can be seen in Figure 21a very clearly, all selected cells either have a higher enrichment score or only positive enrichment scores in the presence of high expression EGFR marker around the CD8+ T cell neighbourhood. Since the filter selected low EGFR expressing cells preferably (Figure 17), those results can be discarded. Thus, the first filter did not learn any useful information about enrichments.

Figure 21d shows the findings of the second CellCnn filter, the high expression selection filter (Figure 18). T regulatory cells, CD68+CD163+ macrophages and plasma cells have large peaks in both high and low EGFR expression plots in Figure 21e and 21f respectively and

corresponding bubbles in the positive quadrant in Figure 21d. These cell types show enrichment for both label expressions, but when comparing the scale in Figure 21e to 21f and observing the bubble graph in Figure 21e, it appears that the enrichment score is higher for the low label expression compared to the high EGFR label expression, suggesting that these cell types are more enriched in the proximity of low EGFR expression around the vicinity of CD8+ T cells. The granulocytes are close to the y axis because the low expression (Figure 21e) shows the Granulocyte bar below the x axis. However, the high expression in Figure 21f shows a positive value for granulocytes with an error bar above the x axis, suggesting a positive enrichment of granulocytes in the vicinity of high EGFR expression. B cells, CD4+CD45+ T cells, and vasculature cells also show positive enrichment only for enrichment scores of the high marker expression.

3.2.3.3.3 CD4+CD45+RO T cells with ICOS

Figure 22 illustrates the bubble plots and bar plots for CD4+CD45RO+ T cells with ICOS label.

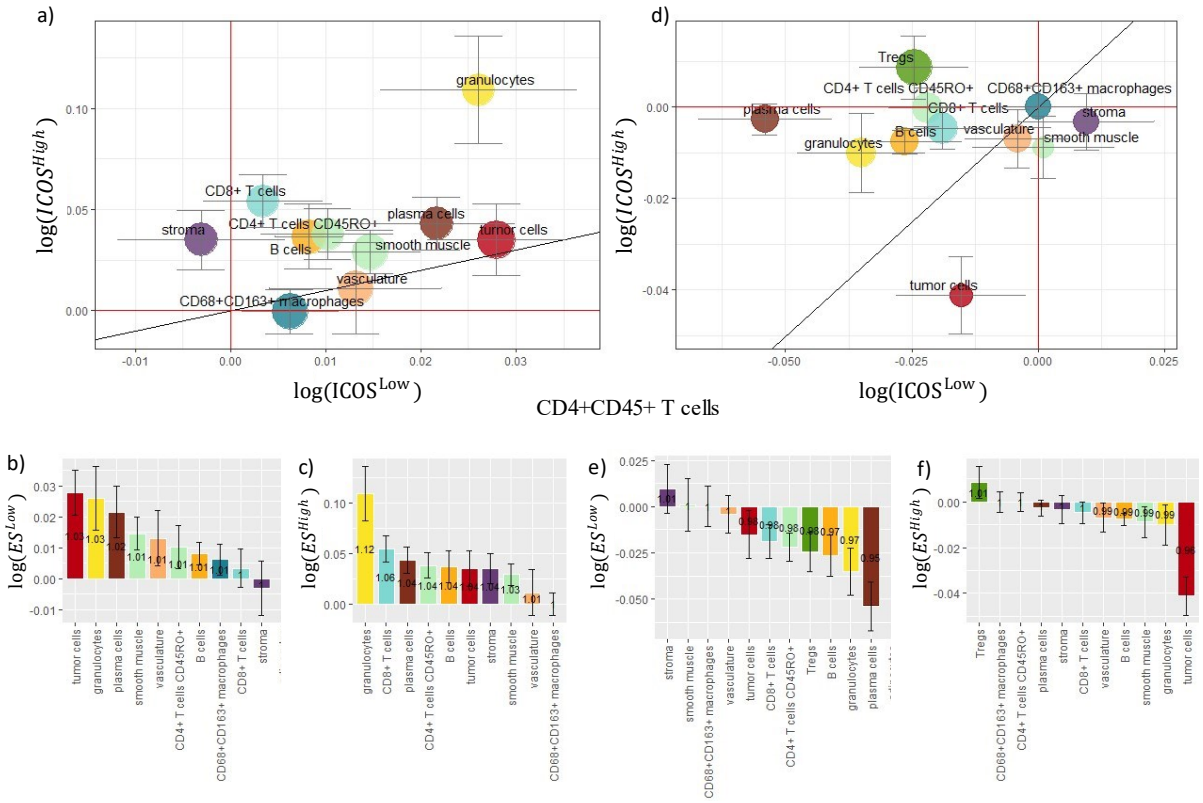


Figure 22 illustrates the logarithm of the enrichment score of low expression plotted against the high ICOS expression for both filters illustrated in Figure 22a and 22d respectively. In Figures 22b and 22c the logarithmic enrichment score of the respective low- and high-label expressions or depicted individually, with the plots in 22e and 22f being the equivalent for the second filter.

Figure 22a, the first filter bubble plot, illustrated a very prominent outlier hinting at the enrichment of granulocytes, present in both high and low label expression, but with a higher value for the high ICOS expression, as can also be seen in Figure 22b when compared to Figure 22c, where granulocytes are the largest intensity bar in the plot. The high ICOS marker expression seen in Figure 22c also shows enrichment in CD8+ T cells, B cells and Cd4+Cd45RO+ T cells, which are all present with positive values for both label expression but with higher log enrichment scores in the high label expression. Similar to before, since the first filter selected cells with low label expression with a higher frequency, those cells will not be considered for further analysis. The bubble plot in Figure 22a, illustrates very well how all other cell types are very close to the trendline or have higher values in the high expression enrichment compared to the low expression. Thus, viable information can be obtained for the first filter.

The right-hand side of Figure 22, the second training filter, shows very poor enrichment scores for both expression groups. As can be further seen in Figure 22e and 22f, most enrichment scores are negative. The only positive results show an enrichment of stroma cells in the proximity of low ICOS expression, which has to be discarded for this filter, since it mostly selected high expression filters (Figure 19). The only visible enrichment can be observed for T regulatory cells, in the high ICOS expression vicinity, in the neighbourhood of the anchor cell.

3.2.4 Ordinal Regression

In the first section of this chapter the best results of the previous section will be compared to their ordinal counterpart. After, the best ordinal result will be analysed. The results obtained were not very good and will be further discussed in the Discussion. However, the Spearman score suggested that some correlation between these results.

The ordinal correlation plot points are usually spread out vertically, thus often not resulting in good R^2 values. Therefore, the Spearman rank correlation coefficient will also be considered to determine whether an underlying correlation may exist.

For the first comparison, granulocytes with 20 nearest neighbours and the EGFR functional marker, was computed ordinally with 5 bins and resulted in Figure 23.

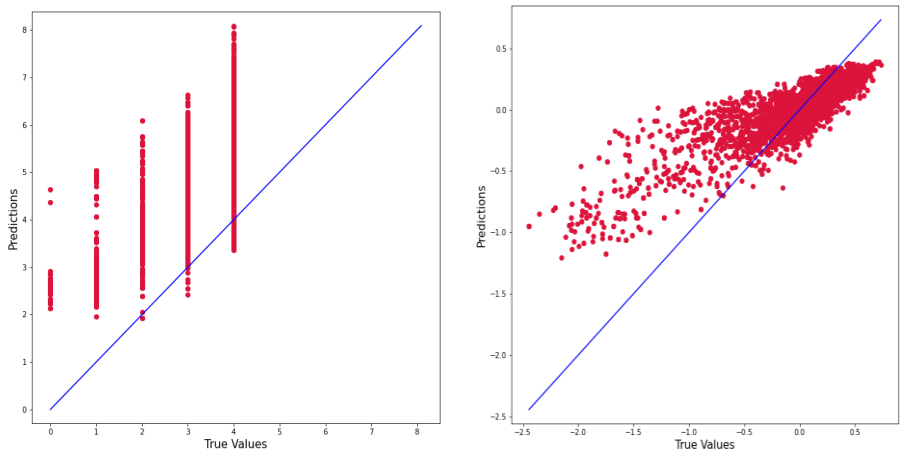


Figure 23 shows the correlation plot for ordinal regression on the left, and the correlation plot for regular regression, for Figure 16b on the right.

The ordinal regression seems to follow the general progression of the trendline but shifted on the y-axis, predicting higher values that are not in the true value set while not predicting values below 2, for data points x, where $0 \geq x \leq 5$. As can be seen in Table 6 below, the spearman's coefficient is between 0 and 1 with 0.67, thus suggesting a moderate positive correlation.

Table 6 shows the metric scores for Granulocyte anchor cell with the functional EGFR marker, comparing ordinal and regular regression option of CellCnn after training.

	R^2 (%)	RMSE	Spearman
Ordinal	-501	1.8	0.67
Regular	63.8	0.26	0.88

For CD8+ T cells, the results of ordinal regression training and the corresponding regular regression are depicted in Figure 24.

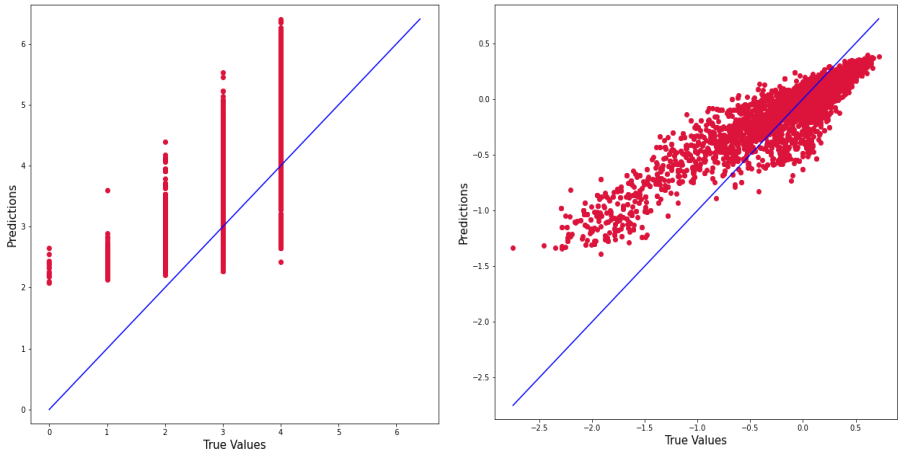


Figure 24 compares the ordinal regression correlation plot of CD8+ T cells in the vicinity of the EGFR label to the regular counterpart, correlation plot of Figure 16d.

The ordinal correlation plot of Figure 24 (left) shows that the progression of the prediction versus the true values appear more along a straight horizontal line, rather than the trendline. However, Table 7 below shows a higher Spearman's rank coefficient than depicted in Table 6 before. Since the R^2 score is negative and the RMSE score by itself is not very informative, no further comments can be given about the accuracy of this training.

Table 7 compared ordinal regression metrics to regular regression metric scores of a similar training setup.

	R^2 (%)	RMSE	Spearman
Ordinal	-43	0.99	0.77
Regular	73.1	0.29	0.89

The final ordinal and regular comparison, refers to CD4+Cd45+RO T cells in the proximity of the ICOS functional marker. As can be seen in Figure 25, the prediction versus the true values in the correlation plot of ordinal regression (left), show that the graph follows the trend line but

shifted on the y-axis as well. Some lower values are predicted again not entirely and on the upper range it predicts outside the range.

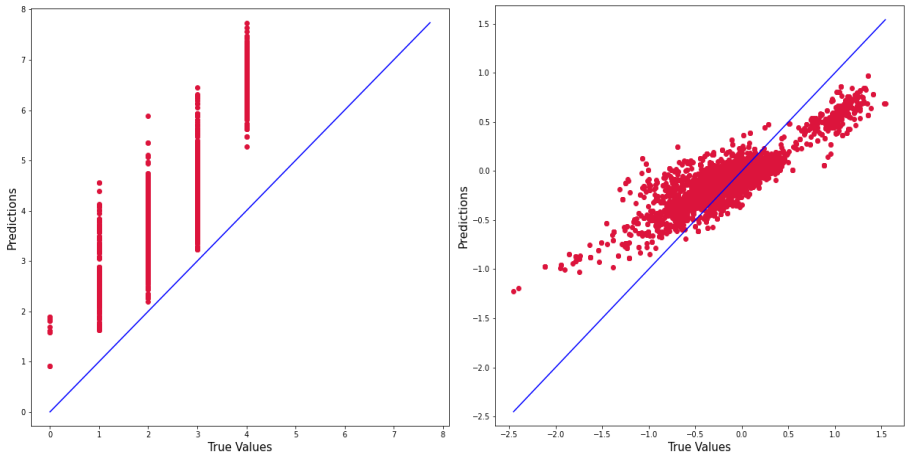


Figure 25 shows the correlation plots for ordinal regression (left) and the corresponding linear regression correlation plot from Figure 16f (right).

As seen in Table 8 below, the R^2 of the ordinal setup is negative and the RMSE score above 1. The Spearman rank correlation coefficient is getting closer to 1, suggesting that some positive correlation exists.

Table 8: Compares the metric scores for ordinal and regular regression of a comparable problem setup.

	R^2 (%)	RMSE	Spearman
Ordinal	-397	1.53	0.70
Regular	65.9	0.27	0.81

The best ordinal regression result came from B cell neighbourhood in the proximity of EGFR with 10 nearest neighbours. The scores are 40% for R^2, 0.62 for RMSE and a Spearman's rank coefficient of 0.77. The correlation plot shown in Figure 26, shows that the true values versus the prediction graph follows the progression of the trend plot fairly well, but also seems to be shifted on the y-axis.

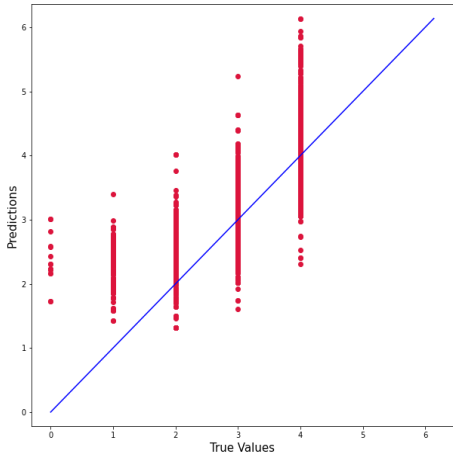


Figure 26 shows the correlation plot for ordinal regression of B cells in the vicinity of EGFR marker expression.

4. Discussion

In the following, the results of this thesis will be discussed and compared with current literature and analysed in the context of their biological relevance.

First, very briefly, the framework testing will be discussed. The optimization process will be analysed briefly after, then the main part of this discussion will be the downstream analysis and related enrichment analysis. Finally, the ordinal regression option of CellCnn will be discussed to conclude the discussion section.

4.1 Testing the Framework

In the linear toy model, a problem structure similar to the main problem has been constructed using artificially created data with similar dimensionality to the main dataset in this thesis. This approach was compared to a TensorFlow linear model, and the two approaches returned similar scores. The Toy model using CellCnn got 97.5% and the TensorFlow 99.89% accuracy using the R^2 score measure. The linear toy model shows, that CellCnn is working at least around what you would expect from a regular neural network type.

In the second part of the framework testing, the frequencies of certain cells referred to as target cells, was computed around an anchor cell and used as label during training of CellCnn.

For the first example, the frequency of all CD4+ T Cell types and subtypes has been computed around the neighbourhood of granulocytes, with the k nearest neighbours being 50.

This example was inspired by the paper of Schürch et al (2020), who found an enrichment of PD-1+ CD4+ T cells in the vicinity of granulocytes. The first example gave the best results of the frequency analysis using both patient groups, DII and CLR. The R^2 score was 52.4%. The other groups individually got much lower scores below 30%. The best score of 52.4% suggest some kind of correlation, but is still not as good as expected, based on other results in this thesis. In the second example the scores got much worse, analysing the frequency of CD68+ macrophages in the vicinity of granulocytes returned R^2 values below 10% for all group combinations, and all RMSE scores above 1. One reason for the poor results might be the amount of k nearest neighbours being set to 50. Due to the computation of a neighbourhood being restricted to cells in the same patient for a given survival group and the TMA regions, the set of available cells per neighbourhood that met those requirements around that given anchor cell was sometimes only around 80 cells. Thus, using much more than 50 cells in any analysis does not make sense since this constitutes more than 50% of some of the available and computed functional cell neighbourhoods. A possible case would be that increasing the

neighbourhoods to 100 cells might have given more accurate result, since we are dealing with frequencies.

4.2 Discussion of Anchor-based functional Spatial enrichment analysis

The anchor-based enrichment analysis is the main focus of this thesis. In the first sub-chapter the optimization will be discussed, which served as a basis for the following downstream analysis of the best results.

4.2.1 Optimizing the training process

Before the main optimization could begin, a starting point needed to be set. For this step, several cell types were used to compute cellular neighbourhoods and compared by using different functional marker for each of those neighbourhoods per cell type used. B cells, regulatory T cells, CD4+ CD45 RO T cells, CD8+ T cells and granulocytes were the selected cell types. The functional markers are explained in Table 1, in chapter 2.1.2.

As can be seen from Table 4, the best results came from the functional marker of ICOS with CD4+CD45RO+ T cells, as well as EGFR with granulocytes and CD8+ T cells. Beta-catenin was further included in the next step but showed R^2 scores below 50%, which does not show a high enough score compared to other results produced during this phase. Binning did not work well with the majority of cases, and generally performed worse than regular regression, as can be seen in Table 5, thus it was excluded from further analysis.

Ki67, the proliferation marker, had the worse result of all selected functional markers. Some training attempts, and even failed for some cell types such as B cells. High expressions of Ki67 can be linked to a higher survival chance in colorectal cancer (Melling et al., 2016; Salminen et al., 2005) according to some authors, is an indication for a low survival chance according to other authors (Luo et al., 2019), and appears to be a controversial topic.

However, several studies identify Ki67 as a prognostic marker for colorectal cancer and different usage strategies have been proposed (Made Mulyawan, 2019) implying some relation with Ki67 expression and survival exists. This was one of the reasons that made Ki67 an interesting functional marker for the first functional marker in this thesis. The resulting poor performance during training is therefore less likely due to a biological reason and might be related to unknown factors in the dataset or during training, simply not providing a good pattern for CellCnn to learn from.

EGFR resulted in some of the best scores across different selected cell types, also returning the best overall score. The good results for the EGFR marker might be attributed to the role EGFR plays in the development of colorectal cancer, specifically. Authors O'dwyer & Benson (2002) reported an expression of EGFR in human colorectal cancer from roughly 65% to 70% of cases (O'dwyer & Benson, 2002), and EGFR seems to be widely recognized as an important factor for colorectal cancer development (Li et al., 2020; Markman et al., 2010; Yarom & Jonker, 2011). The resulting occurrence of EGFR marker intensities across colorectal cancer patients might provide a good learning platform for CellCnn to train from.
CD8+ T cells reported the highest overall score with EGFR, as can be seen in Figure 8, Figure 13 and also Figure 16c. In non-small cell lung cancer (NSCLC), EGFR was reported by Nishii et al. (2022) to have a tumour microenvironment with low CD8+ T cell expression and inhibiting EGFR increased the overall CD8+ T cell expression (Nishii et al., 2022; Selenz et al., 2022). It is unclear how this case of NSCLC can be generalized and would translate to colorectal cancer. However, this might suggest an absence of CD8+ T cells in a tumour microenvironment with high EGFR expression. This could be further explained by the discussion in section 4.2.2.1 and findings of section 3.2.3.3.1, where CD8+ T cells were only found in the vicinity of granulocytes with a low EGFR expression. Thus, there seems to be a biological relation between CD8+ T cells and EGFR, but it is unclear how this affected the results of CellCnn, if at all. Although, a potential absence of CD8+ T cells around EGFR expression does not give any hints about the actual neighbourhood around those cells per se, which would be evaluated during training of CellCnn.

As can be seen in Figure 9, the highest score for BCL-2, the apoptosis marker (Table 1) was below 50% for the R^2 score, with CD8+ T cells using binned values, and only 33.11% R^2score for B cells with regular regression. Since binning was not further pursued and ~33% is generally not a good score for correlation, BCL-2 was not further considered as a functional marker in this work. However, it may be interesting to note that B cells had the highest result for BCL-2, which as mentioned in Table 1, plays a key role in the regulation and development of B cells (Adams et al., 2018; Kale et al., 2018b). This is very consistent with results from the EGFR functional marker and also the following marker ICOS, discusses in the following part. CellCnn seems to get the highest results, and therefore learn the most patterns, for biologically relevant functional marker and cell type pairs.

CD4+CD45RO+ cells performed best with the ICOS functional marker, out of the chosen cell types for testing. CD4 T cells that express CD45RO antigens are called memory T cells, they grow after being exposed to memorised recall antigens and can differentiate from naïve T cells (Devi et al., 2017). ICOS has a very high expression on memory T cells, as costimulation with ICOS increases proliferation and aids in survival of CD4+ T cells (Moore et al., 2011).
Thus, the results being better with ICOS compared to other functional markers as observed in Figure 11, in the proximity of CD4+CD45RO cells could be explained by the connection between ICOS and CD4+ memory T cells that express CD45RO. As ICOS has a high expression around those memory T cells, many cells could have been selected for the set of available neighbours creating a higher variety for CellCnn to learn from. This might be further explained by the next highest result seen in Figure 11, CD8+ T cells, which similar to CD4+ T cells can express the memory effector CD45RO, and also shows strong ICOS expressions (Machura et al., 2008; Nagamatsu et al., 2011), which might explain why those two cell types performed the best out of the selected cell types for training.

After deciding on the starting cell types, the best results up to that point were further improved. CD4+CD45RO+ T cells with ICOS, granulocytes with EGFR and beta-catenin and CD8+ T cells with EGFR functional marker were selected for further improvement, by changing the computation of the label and number of markers, additionally comparing 48 and 56. Between the different label computations average, maximum, minimum, median and percentile of minimum or maximum. Average performed the best across most cases and using all 56 markers consistently performed better than 48 markers, with respect to the R^2 score.

In the last optimization step, a trend plot was produced, by increasing the k-nearest neighbourhood, there is no reason than could explain whether a smaller or larger neighbourhood might train better or worse for CellCnn, thus this was more of a heuristic approach for optimizing this hyperparameter and the overall accuracy. Using 20 nearest neighbours (k=20) performed best for all 3 cases, within a reasonable margin of error.

4.2.2 Downstream Analysis

4.2.2.1 Granulocytes with EGFR marker

As can be inferred from Table 1, the epidermal growth factor receptor (EGFR) can be involved in various biological processes related to cancer. The role of EGFR for cancer prognosis is controversial, but it is often believed that the presence of EGFR might be linked to a poor prognosis in cases such as colorectal cancer patients and is overexpressed for those cancer types (Nicholson et al., 2001; Spano et al., 2005). For the first filter, which selected cells showing a low expression of EGFR in the proximity of granulocytes more frequently, an enrichment in CD8+ T cells, adipocytes, and plasma cells was detected. The second filter selected cells with a high marker expression more frequently and detected an enrichments of B cells in the neighbourhood around granulocytes with a high EGFR expression compared to the low expression. The paper in the appendix by Babaei et al. (Appendix I) also features these results and found that the CD8+ T cell enrichment can be linked to "significantly reduced cytokine expression and proliferation activity", due to the inhibitory role of granulocytes with respect to cytokine production by T cells, for the first filter. The second filter, which selected high EGFR expression cells more prominently, explained that B cells and granulocytes can be linked to high EGFR expression singling in the tumour microenvironment, which was reported to support previous findings which suggested granulocytes as a contributor in differentiation of neoplastic B cells (Babaei et al., n.d., Appendix I).

4.2.2.2 CD8+ T cells with EGFR marker and k=20

Figure 18 shows the filter selection for the EGFR marker in the vicinity of CD8+ T cells, where the first filter selected low EGFR expressions more frequently than the second filter and vice versa. The first filter did not detect any enrichments that can be considered for further discussion. However, the second filter found an enrichment of granulocytes, B cells and CD4+CD45+ T cells in proximity of high EGFR expression around CD8+ T cells. This result might be explained by looking at the results discussed in 4.2.2.1, where granulocytes had CD8+ T cells in the low EGFR expression and B cells around high EGFR expression. The enrichment of B cells could be attributed to the enrichment of granulocytes, as explained above, there exists evidence of a connection between granulocytes and B cells since granulocytes play a role in differentiation a subtype of B cells, both linked to a high EGFR expression. The last enrichment detected CD4+CD45+ T cells. CD45, has a target receptor on the CD22 molecules of B cells, and CD45 expression can be upregulated during B cell differentiation and maintains its stability

even on mature B cells (Hendrickx & Bossuyt, 2001; Ledbetter et al., 1993, p. 8). Since an enrichment of B cells was found, the CD4+CD45+ T cell enrichment might have been affected by that.

4.2.2.3 CD45-CD45RO with the ICOS marker and k=20

Filter selection for the ICOS marker in the vicinity of Cd4+CD45RO+ T cells, selected cells with a low ICOS expression in the first filter and high ICOS expression in the second filter. The first filter found no significant enrichment for any cell types. The second filter found an enrichment of T Regulatory cells in the vicinity of CD4+CD45RO+ cells around a high ICOS expression. Burmeister et al. (2018) found a high ICOS expression on T Regulatory cells, specifically FoxP3+ T Reg cells, as well as CD4+ based Memory T cells in a mice model. The authors suggested that ICOS may be involved in regulation of both T Regulatory cells and Memory T cells derived from CD4+ T cells (Burmeister et al., 2008). These findings are further supported by the paper from Chen et al. (2018), who discovered the role of ICOS in upregulating the suppressive ability of T Regulatory cells, by inducing transcription of FoxP3 (Chen et al., 2018).

Faget et al. (2012) reported a connection with Memory CD4+ T cells and T regulatory cells in human breast cancer. The authors found that T Regulatory cells in the tumour microenvironment expressed ICOS, which played a major role in suppressing Memory CD4+ T cells in the tumour microenvironment (Faget et al., 2012). These findings may explain the enrichment of T Regulatory cells in the proximity of CD4+CD45RO+ cells (Memory T cells) with a high ICOS expression.

4.3 Ordinal Regression

For the ordinal regression problem, first, the three best result from the optimization of the regular regression were compared to their ordinal counterpart. The regular regression and ordinal regression computed 20 k-nearest neighbours with the ordinal regression having 5 bins for each case respectively. For granulocytes and the EGFR marker, the correlation plot in Figure 23, seemed to follow the general progression of the trendline but shifted on the y-axis, while the Spearman's ran coefficient showed a score 0.67 for ordinal versus 0.88 for regular regression in Table 6. Ordinal regression of CD8+ T cells with the EGFR marker, showed a more horizontal line for the correlation plot (Figure 24), but a better Spearman's rank corelation coefficient with 0.77, while the regular regression was comparable to the result before (0.89), seen in Table 7. The final ordinal and regular comparison, CD4+CD45RO+ cells with the ICOS functional marker again showed a shifted correlation plot for the ordinal regression, seen in Figure 25.

The Spearman's rank correlation of ordinal regression, 0.7 compared to the same metric for regular being 0.81 (Table 8), revealed that all three cases show some correlation with the ordinal regression, but all had poor performance with respect to both R^2 score and RMSE scores.

The best ordinal regression, B cells with the EGFR functional marker 10 nearest neighbours and 5 bins, gave metric scores of 40% and 0.62 for R^2 and RMSE respectively, and a Spearman's rank coefficient of 0.77. Although the Spearman's rank coefficient was not sufficiently higher compared to the previous cases, both the R^2 and RMSE scores were much better. In terms of accuracy, 40% for the R^2 cannot be considered a good performance, but together with the Spearman's rank correlation coefficient, hints at a connecting that might become subject of future research.

5. Conclusion

First the general predictive power of CellCnn was tested, and CellCnn demonstrated in the linear regression model a similar result to the TensorFlow linear neural network. To increase the complexity and thematic relevance, frequency analysis of target cell types around an anchor cell were computed and tested. In the first frequency analysis, one of the key findings of the paper from Schürch et al (2020) derived the first problem tested within the frequency analysis, the frequency of CD4+ T cells in the proximity of granulocytes. The result was the highest metric scores among tested frequency analysis cases that were tested, but still not as high as some of the other metric scores in this thesis. The constraints on the viable neighbourhood size could have been one reason for that.

For the following enrichment analysis, CellCnn seems to have returned the highest correlation accuracies, with respect to R^2 and RMSE scores, for biologically relevant functional marker and cell type pairs, as discussed in chapter 4.2.1. The anchor-based functional spatial enrichment analysis found an enrichment of CD8+ T cells in low expression EGFR neighbourhoods of granulocytes, B cells in high expression EGFR proximity, among other results. CD4+CD45+ T cell, B cell and granulocyte enrichments were found in the high EGFR expression proximity of CD8+ T cells and T Regulatory cells in the proximity of high ICOS expression of CD4+CD45RO+ T cells. Research suggested some explanations for the found enrichments, however more in-depth research will be required to test and verify these findings. They propose hypothesis that aligns with some findings presented in section 4.2.2, derived from current research, and may provide interesting options for future research, potential treatment options and understanding of colorectal cancer. As stated before, the immune tumour microenvironment (ITME) plays a fundamental role in the progression and control of cancer, thus further research, and verification of the found enrichments in this thesis may provide new insight into strategies – toward the ITME in colorectal cancer patients. Examples might include the inhibition of EGFR in the proximity of granulocytes to upregulate CD8+ T cells or perhaps the relation of ICOS to regulatory T cells in human breast cancer, if comparable connections can be clinically proven in future papers.

The ordinal regression setting of CellCnn did not produce high R^2 and RMSE scores but showed some correlations according to the Spearman rank correlation coefficient. Due to the time restriction for this master thesis, ordinal regression could not be further pursued or enhanced, however implementing a refined version of ordinal regression with CellCnn or a derivation of CellCnn, might prove to be an interesting alternative to the classical regression option explored in this thesis.

References

Adams, C. M., Clark-Garvey, S., Porcu, P., & Eischen, C. M. (2018). Targeting the Bcl-2 Family in B Cell Lymphoma. *Front. Oncol.*, *8*, 636.

Agostino, M. D., & Dardanoni, V. (2009). What's so special about Euclidean distance? A characterization with applications to mobility and spatial voting. *Soc. Choice Welfare*, *33*, 211–233.

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *J Big Data*, *8*(1), 53.

Arvaniti, E., & Claassen, M. (2017). Sensitive detection of rare disease-associated cell subsets via representation learning. *Nat. Commun.*, *8*, 14825.

Babaei, S., Christ, J., Makky, A., Zidane, M., Schürch, M., , Christian, & Claasen, M. (n.d.). *S3-CIMA: Supervised spatial single-cell image analysis for the identification of disease-associated cell type compositions in tissue*.

Bae, S., Choi, H., & Lee, D. S. (2021). Discovery of molecular features underlying the morphological landscape by integrating spatial transcriptomic data with deep features of tissue images. *Nucleic Acids Res.*, *49*(10), e55.

Biscione, V., & Bowers, J. (2020). *Learning Translation Invariance in CNNs*.

Black, S., Phillips, D., Hickey, J. W., Kennedy-Darling, J., Venkataraaman, V. G., Samusik, N., Goltsev, Y., Schürch, C. M., & Nolan, G. P. (2021). CODEX multiplexed tissue imaging with DNA-conjugated antibodies. *Nat. Protoc.*, *16*(8), 3802–3835.

Brown, K. W. (2021). *R Gallery Book*. https://bookdown.org/content/b298e479-b1ab-49fa-b83d-a57c2b034d49/

Burmeister, Y., Lischke, T., Dahler, A. C., Mages, H. W., Lam, K.-P., Coyle, A. J., Kroczek, R. A., & Hutloff, A. (2008). ICOS controls the pool size of effector-memory and regulatory T cells. *J. Immunol.*, *180*(2), 774–782.

Carbonneau, M.-A., Cheplygina, V., Granger, E., & Gagnon, G. (2016). *Multiple Instance Learning: A Survey of Problem Characteristics and Applications*.

Cazzato, G., Caporusso, C., Arezzo, F., Cimmino, A., Colagrande, A., Loizzi, V., Cormio, G., Lettini, T., Maiorano, E., Scarcella, V. S., Tarantino, P., Marrone, M., Stellacci, A., Parente, P., Romita, P., De Marco, A., Venerito, V., Foti, C., Ingravallo, G., … Resta, L. (2021). Formalin-Fixed and Paraffin-Embedded Samples for Next Generation Sequencing: Problems and Solutions. *Genes*, *12*(10).

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?– Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Chen, Q., Mo, L., Cai, X., Wei, L., Xie, Z., Li, H., Li, J., & Hu, Z. (2018). ICOS signal facilitates Foxp3 transcription to favor suppressive function of regulatory T cells. *Int. J. Med. Sci.*, *15*(7), 666–673.

Chew, V., Toh, H. C., & Abastado, J.-P. (2012). Immune microenvironment in tumor progression: Characteristics and challenges for therapy. *J. Oncol.*, *2012*, 608406.

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Comput Sci*, *7*, e623.

Colbeck, E. J., Ager, A., Gallimore, A., & Jones, G. W. (2017). *Tertiary Lymphoid Structures in Cancer: Drivers of Antitumor Immunity, Immunosuppression, or Bystander Sentinels in Disease?*

Devi, M., Vijayalakshmi, D., Dhivya, K., & Janane, M. (2017). Memory T Cells (CD45RO) Role and Evaluation in Pathogenesis of Lichen Planus and Lichenoid Mucositis. *J. Clin. Diagn. Res.*, *11*(5), ZC84–ZC86.

Dougherty, J., & Ilyankou, I. (2022). *Hands-On Data Visualization*. https://handsondataviz.org/

Faget, J., Bendriss-Vermare, N., Gobert, M., Durand, I., Olive, D., Biota, C., Bachelot, T., Treilleux, I., Goddard-Leon, S., Lavergne, E., Chabaud, S., Blay, J. Y., Caux, C., & Ménétrier-Caux, C. (2012). ICOS-ligand expression on plasmacytoid dendritic cells supports breast cancer progression by promoting the accumulation of immunosuppressive CD4+ T cells. *Cancer Res.*, *72*(23), 6130–6141.

Fang, S., & Wang, Z. (2014). EGFR mutations as a prognostic and predictive marker in non-small-cell lung cancer. *Drug Des. Devel. Ther.*, *8*, 1595–1611.

Filho, D. B. F., J, Júnior, Silva, & Rocha, E. C. (2011). What is R2 all about? *Leviathan*, *3*(3), 60–68.

Gohil, S. H., Iorgulescu, J. B., Braun, D. A., Keskin, D. B., & Livak, K. J. (2021). Applying high-dimensional single-cell technologies to the analysis of cancer immunotherapy. *Nat. Rev. Clin. Oncol.*, *18*(4), 244–256.

Gutiérrez, P. A., Pérez-Ortiz, M., Sánchez-Monedero, J., Fernandez-Navarro, F., & Martínez, C. (2015). Ordinal Regression Methods: Survey and Experimental Study. *IEEE Transactions on Knowledge and Data Engineering*, *28*. https://doi.org/10.1109/TKDE.2015.2457911

Harold, F. M. (2005). Molecules into cells: Specifying spatial architecture. *Microbiol. Mol. Biol. Rev.*, *69*(4), 544–564.

Hendrickx, A., & Bossuyt, X. (2001). Quantification of the leukocyte common antigen (CD45) in mature B-cell malignancies. *Cytometry*, *46*(6), 336–339.

Hochreiter, S. (1998). The Vanishing Gradient Problem During Learning Recurrent Neural Nets and Problem Solutions. *Int. J. Uncertainty Fuzziness Knowledge Based Syst.*, *06*(02), 107–116.

Jansen, C. S., Prokhnevska, N., & Kissick, H. T. (2019). The requirement for immune infiltration and organization in the tumor microenvironment for successful immunotherapy in prostate cancer. *Urol. Oncol.*, *37*(8), 543–555.

Kale, J., Osterlund, E. J., & Andrews, D. W. (2018). BCL-2 family proteins: Changing partners in the dance towards death. *Cell Death Differ.*, *25*(1), 65–80.

Kennedy-Darling, J., Bhate, S. S., Hickey, J. W., Black, S., Barlow, G. L., Vazquez, G., Venkataraaman, V. G., Samusik, N., Goltsev, Y., Schürch, C. M., & Nolan, G. P. (2021). Highly multiplexed tissue imaging using repeated oligonucleotide exchange reaction. *Eur. J. Immunol.*, *51*(5), 1262–1277.

Labani-Motlagh, A., Ashja-Mahdavi, M., & Loskog, A. (2020). The Tumor Microenvironment: A Milieu Hindering and Obstructing Antitumor Immune Responses. *Front. Immunol.*, *11*, 940.

Ledbetter, J. A., Deans, J. P., Aruffo, A., Grosmaire, L. S., Kanner, S. B., Bolen, J. B., & Schieven, G. L. (1993). CD4, CD8 and the role of CD45 in T-cell activation. *Curr. Opin. Immunol.*, *5*(3), 334–340.

Li, Q.-H., Wang, Y.-Z., Tu, J., Liu, C.-W., Yuan, Y.-J., Lin, R., He, W.-L., Cai, S.-R., He, Y.-L., & Ye, J.-N. (2020). Anti-EGFR therapy in metastatic colorectal cancer: Mechanisms and potential regimens of drug resistance. *Gastroenterol. Rep.*, *8*(3), 179–191.

Lindsay, G. W. (2021). Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. *J. Cogn. Neurosci.*, *33*(10), 2017–2031.

Lipton, Z. C. (2015). *A Critical Review of Recurrent Neural Networks for Sequence Learning*.

Luo, Z.-W., Zhu, M.-G., Zhang, Z.-Q., Ye, F.-J., Huang, W.-H., & Luo, X.-Z. (2019). Increased expression of Ki-67 is a poor prognostic marker for colorectal cancer patients: A meta analysis. *BMC Cancer*, *19*(1), 123.

Machura, E., Mazur, B., Pieniazek, W., & Karczewska, K. (2008). Expression of naive/memory (CD45RA/CD45RO) markers by peripheral blood CD4+ and CD8 + T cells in children with asthma. *Arch. Immunol. Ther. Exp.*, *56*(1), 55–62.

Made Mulyawan, I. (2019). Role of Ki67 protein in colorectal cancer. *International Journal of Research in Medical Sciences*, *7*(2), 644–648.

Markman, B., Javier Ramos, F., Capdevila, J., & Tabernero, J. (2010). EGFR and KRAS in colorectal cancer. *Adv. Clin. Chem.*, *51*, 71–119.

Melling, N., Kowitz, C. M., Simon, R., Bokemeyer, C., Terracciano, L., Sauter, G., Izbicki, J. R., & Marx, A. H. (2016). High Ki67 expression is an independent good prognostic marker in colorectal cancer. *J. Clin. Pathol.*, *69*(3), 209–214.

Moore, T. V., Clay, B. S., Ferreira, C. M., Williams, J. W., Rogozinska, M., Cannon, J. L., Shilling, R. A., Marzo, A. L., & Sperling, A. I. (2011). Protective effector memory CD4 T cells depend on ICOS for survival. *PLoS One*, *6*(2), e16529.

Mukaka, M. M. (2012). Statistics corner: A guide to appropriate use of correlation coefficient in medical research. *Malawi Med. J.*, *24*(3), 69–71.

Nagamatsu, T., Barrier, B. F., & Schust, D. J. (2011). The regulation of T-cell cytokine production by ICOS-B7H2 interactions at the human fetomaternal interface. *Immunol. Cell Biol.*, *89*(3), 417–425.

Namatēvs, I. (2017). Deep Convolutional Neural Networks: Structure, Feature Extraction and Training. *Information Technology and Management Science*, *20*(1).

Nicholson, R. I., Gee, J. M., & Harper, M. E. (2001). EGFR and cancer prognosis. *Eur. J. Cancer*, *37 Suppl 4*, S9-15.

Nishii, K., Ohashi, K., Tomida, S., Nakasuka, T., Hirabae, A., Okawa, S., Nishimura, J., Higo, H., Watanabe, H., Kano, H., Ando, C., Makimoto, G., Ninomiya, K., Kato, Y., Kubo, T., Ichihara, E., Hotta, K., Tabata, M., Toyooka, S., … Kiura, K. (2022). CD8+ T-cell Responses Are Boosted by Dual PD-1/VEGFR2 Blockade after EGFR Inhibition in Egfr-Mutant Lung Cancer. *Cancer Immunol Res*, *10*(9), 1111–1126.

O'dwyer, P. J., & Benson, A. B., 3rd. (2002). Epidermal growth factor receptor-targeted therapy in colorectal cancer. *Semin. Oncol.*, *29*(5 Suppl 14), 10–17.

O'Shea, K., & Nash, R. (2015). *An Introduction to Convolutional Neural Networks*.

Oyeka, I. C. A., & Ebuh, G. U. (2012). Modified Wilcoxon Signed-Rank Test. *Open Journal of Statistics*, *02*(02), 172–176.

Phillips, D., Schürch, C. M., Khodadoust, M. S., Kim, Y. H., Nolan, G. P., & Jiang, S. (2021). Highly Multiplexed Phenotyping of Immunoregulatory Proteins in the Tumor Microenvironment by CODEX Tissue Imaging. *Front. Immunol.*, *12*, 687673.

Rehmer, A., & Kroll, A. (2020). On the vanishing and exploding gradient problem in Gated Recurrent Units. *IFAC-PapersOnLine*, *53*(2), 1243–1248.

Salminen, E., Palmu, S., Vahlberg, T., Roberts, P.-J., & Söderström, K.-O. (2005). Increased proliferation activity measured by immunoreactive Ki67 is associated with survival improvement in rectal/recto sigmoid cancer. *World J. Gastroenterol.*, *11*(21), 3245–3249.

Schürch, C. M., Bhate, S. S., Barlow, G. L., Phillips, D. J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D. R., Kinoshita, S., Samusik, N., Goltsev, Y., & Nolan, G. P. (2020). Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell*, *183*(3), 838.

Selenz, C., Compes, A., Nill, M., Borchmann, S., Odenthal, M., Florin, A., Brägelmann, J., Büttner, R., Meder, L., & Ullrich, R. T. (2022). EGFR Inhibition Strongly Modulates the Tumour Immune Microenvironment in EGFR-Driven Non-Small-Cell Lung Cancer. *Cancers*, *14*(16).

Sharkawy, A.-N. (2020). Principle of Neural Network and Its Main Types: Review. *Journal of Advances in Applied & Computational Mathematics*, *7*(1), 8–19.

Sobecki, M., Mrouj, K., Camasses, A., Parisis, N., Nicolas, E., Llères, D., Gerbe, F., Prieto, S., Krasinska, L., David, A., Eguren, M., Birling, M.-C., Urbach, S., Hem, S., Déjardin, J., Malumbres, M., Jay, P., Dulic, V., Lafontaine, D. L., … Fisher, D. (2016). The cell proliferation antigen Ki-67 organises heterochromatin. *Elife*, *5*, e13722.

Spano, J.-P., Lagorce, C., Atlan, D., Milano, G., Domont, J., Benamouzig, R., Attar, A., Benichou, J., Martin, A., Morere, J.-F., Raphael, M., Penault-Llorca, F., Breau, J.-L., Fagard, R., Khayat, D., & Wind, P. (2005). Impact of EGFR expression on colorectal cancer patient prognosis and survival. *Ann. Oncol.*, *16*(1), 102–108.

Tarnate, K. J. (2020). Overcoming the vanishing gradient problem of Recurrent Neural Networks in the Iso 9001 quality management audit reports classification. *INTERNATIONAL JOURNAL OF SCIENTIFIC & TECHNOLOGY RESEARCH*, *9*(03).

Trajkovski, G., Ognjenovic, L., Karadzov, Z., Jota, G., Hadzi-Manchev, D., Kostovski, O., Volcevski, G., Trajkovska, V., Nikolova, D., Spasevska, L., Janevska, V., & Janevski, V. (2018). Tertiary Lymphoid Structures in Colorectal Cancers and Their Prognostic Value. *Open Access Maced J Med Sci*, *6*(10), 1824–1828.

Valenta, T., Hausmann, G., & Basler, K. (2012). The many faces and functions of β-catenin. *EMBO J.*, *31*(12), 2714–2736.

Wang, B., Li, X., Liu, L., & Wang, M. (2020). β-Catenin: Oncogenic role and therapeutic target in cervical cancer. *Biol. Res.*, *53*(1), 33.

Watada, J. (2010). *Kolmogorov-Smirnov Two Sample Test with Continuous Fuzzy Data*. 175–186.

Xiao, Z., Mayer, A. T., Nobashi, T. W., & Gambhir, S. S. (2020). ICOS Is an Indicator of T-cell-Mediated Response to Cancer Immunotherapy. *Cancer Res.*, *80*(14), 3023–3032.

Yarom, N., & Jonker, D. J. (2011). The role of the epidermal growth factor receptor in the mechanism and treatment of colorectal cancer. *Discov. Med.*, *11*(57), 95–105.

Zhang, X., Nie, D., & Chakrabarty, S. (2010). Growth factors in tumor microenvironment. *Front. Biosci.*, *15*(1), 151–165.

Appendix I

S^3-CIMA: Supervised spatial single-cell image analysis for the identification of disease-associated cell type compositions in tissue

Sepideh Babaei[1], Jonathan Christ[2], Ahmad Makky[3], Mohammed Zidane[3], Christian M. Schürch[3], Manfred Claassen[1,4]

[1]Department of Internal Medicine I, University Hospital Tübingen, Tübingen, Germany
[2]Department of Physics, University of Vienna, Vienna, Austria
[3]Department of Pathology and Neuropathology, University Hospital and Comprehensive Cancer Center Tübingen, Tübingen, Germany
[4]Department of Computer Science, University of Tübingen, Tübingen, Germany

ABSTRACT

The tissue microenvironment, and the spatial organization of cells within it, constitute factors strongly associated with physiological and pathological processes including cancer and autoimmune diseases. Here, we present S^3-CIMA, a weakly supervised convolutional neural network model that enables the detection of disease-specific microenvironment compositions from high-dimensional proteomic imaging data. We demonstrate the utility of S^3-CIMA by determining cancer outcome- and cellular signaling-specific spatial cell state compositions in highly multiplexed fluorescence microscopy data of the tumor microenvironment in colorectal cancer. Moreover, we use S^3-CIMA to identify disease onset-specific changes of the pancreatic tissue microenvironment in type 1 diabetes using imaging mass cytometry data. We expect S^3-CIMA to be a powerful tool to discover novel disease-associated spatial cellular interactions from currently available and future spatial biology datasets.

INTRODUCTION

The tissue microenvironment (TME) constitutes a spatially organized, dynamic and complex system of multiple cell types conferring emergent tissue properties in health and disease. Lymph nodes constitute a seminal example of how spatial organization of cell type composition in the TME confers emergent properties in terms of orchestrating the immune response to pathogens (Gerner et al., 2012). In particular, dendritic cells and T cells are spatially matched in the lymph node's T cell zone to increase the efficiency of adaptive immunity (Qi et al., 2014;

Baptista et al., 2019). Until recently, it has been difficult to study such structures due to the technical requirement to both resolve spatial as well as high dimensional molecular profiles at the single-cell level. Moreover, the subsequent requirements to - typically computationally - integrate and interpret the resulting data to the end of defining descriptions or models of the TME and their disease-associated aberrations have been challenging.

This issue has been addressed technically with the development of high-dimensional proteomic, transcriptomic and epigenomic methods (Thornton et al., 2021, Gizzi et al., 2021), single-cell spatial biology technologies, covering multiplexed fluorescence *in situ* hybridization (FISH), imaging mass cytometry (IMC) (Giesen et al., 2014) and multiplexed microscopy such as cyclic immunofluorescence (CyCIF) (Lin et al., 2022) and co-detection by indexing (CODEX) (Goltsev 2018, Black et al., 2021). These approaches enable joint measurement of dozens of features in FISH, up to hundreds in multiplexed microscopy, and tens of thousands in spot-based spatial transcriptomics. Different resolutions, from local tissue spots as in spatial transcriptomics (e.g., high-definition spatial transcriptomics (Vickovic et al. 2019), Visium (Salmén et al. 2018) and DBit-seq (Liu et al. 2020)) to subcellular resolution (e.g., FISH, IMC, CyCif, CODEX) are possible. Further, dedicated image processing, segmentation, and registration approaches have been developed to process and visualize the resulting data and quantify their respective signal intensities and distributions (Palla et al. 2022).

The resulting data lends itself for defining and mapping out TME arrangements. Two types of TME definitions can be distinguished: (1) **unsupervised** definition of TME arrangements from the spatial single-cell data alone and (2) **supervised** definition of TME characteristics associated with external cues such as disease state. While the first kind of approaches allow for defining cell type composition of the TME, the second kind of approaches go a step further to identify differential cell type composition across conditions, including possibly novel, so far unappreciated cell subtypes. **So far only unsupervised approaches of the first kind have been reported and supervised approaches of the second kind are lacking.**

Approaches of the first kind include “Histology Topography Cytometry Analysis Toolbox” (histoCAT) (Schapiro et al. 2017) enabling unsupervised spatial interrogation of cell–cell interactions in IMC data. The neighborhood analysis in histoCAT examines if a certain cell

type is located significantly closer to another cell type than expected by chance using two individual one-tailed permutation tests. “stLearn” (Spatial Transcriptomics Learn) was developed to calculate cell-cell interactions based on the unsupervised clustering of morphological similarity of spatial transcriptomics data (Pham et al. 2020). Briefly, stLearn utilizes unsupervised clustering to group similar spots into clusters (i.e., spatial morphological gene expression) and then significant cell-cell interactions are detected by a permutation approach. The tissue location hotspots are determined as locations in the tissue where there is both high interaction activity and diverse cell type co-localization. The “Giotto” analyzer and viewer (Dries et al. 2021) consists of tools to process and visualize spatial transcriptomics and proteomics expression data. The toolbox introduced new methods to identify a feature (gene or protein) that constructs a coherent spatial pattern based on unsupervised clustering and statistical enrichment of spatial network neighbors. In sum, these approaches often utilize nearest neighbors and other statistical approaches to identify relationships between different cell types and their TME based on an unsupervised strategy (Palla et al., 2022, Schürch et al., 2020).

The above-mentioned methods are TME inference approaches of the first kind - as defined above - i.e., they are all unsupervised learning approaches that allow to define TME prototypes not allowing to infer TME differences associated and possibly explaining external cues such as disease state. Supervised learning approaches are required to overcome this limitation. Such approaches allow for identifying differences of TME models in comparative study settings, aiming at identifying differences of the same TME prototype across conditions, such as disease state or patient outcome. Supervised learning approaches enabling the detection of disease-associated TME compositions are so far lacking.

To close this gap, we present **supervised spatial single-cell image analysis (S^3-CIMA)** for the identification of disease-associated cell type compositions in tissue microenvironments. S^3-CIMA is a weakly supervised approach that leverages a single-layer convolutional neural network (CNN) architecture (LeCun et al., 2015). We demonstrate the utility of S^3-CIMA by identifying outcome-specific cell state compositions from a CODEX-based study of the colorectal cancer (CRC) TME. Specifically, we applied the method to a cohort of CRC patients

to identify high-risk-specific spatial neighborhood cell type compositions in the tumor (Schürch et al., 2020). Further, we use S^3-CIMA to identify disease onset-specific changes of the pancreatic TME in type 1 diabetes in an IMC study (Damond et al. 2019). We expect that S^3-CIMA will enable the study of other diseases and spatial biology data types and will generally be valuable to identify novel disease-associated cellular interactions.

RESULTS

Weakly supervised learning of disease-associated TME compositions with S^3-CIMA

We aim at learning (disease) condition-associated TME composition from multiparametric and spatially resolved single-cell data. Specifically, we consider the notion of condition-associated TME compositions as the spatial local enrichment of specific cell subsets with respect to a condition, i.e., *supervised spatial enrichment analysis*.

We distinguish three different types of *supervised spatial enrichment analysis.* In *global spatial enrichment analysis,* we aim at identifying which cell subsets are enriched in spatial neighborhoods across conditions, e.g., in CRC tumor tissues of disease manifestations with varying outcomes. In *local (anchor based) spatial enrichment analysis* we instead aim at identifying condition-specific cell subset enrichment in the proximity of an *anchor cell type*, e.g., tumor-infiltrating cytotoxic T cells in CRC with varying outcomes. In *functional spatial enrichment analysis,* we aim at identifying cell subsets enriched in the proximity of a specific functional activity of the tissue, e.g., local signaling activity of cells.

The input of S^3-CIMA is derived from high-dimensional *in situ* proteomic imaging data that has been processed up to single-cell segmentation, giving rise to a data matrix where every single cell (row) is associated with a profile of protein marker expression levels and spatial coordinates in the tissue image (columns) (**Fig. 1a**). The input for S^3-CIMA are tuples of profiles of a set of spatially proximal cells, i.e., k-nearest *cell neighborhoods* (*k-NN*) around an anchor position (e.g., position of a cytotoxic T cell), associated with a phenotype indicating the originating condition (e.g., membership to a sample associated with a specific outcome) (**Fig. 1a**). The association between these cell neighborhoods and their phenotype might be conferred

by the occurrence of a cell subset not matching to a canonical cell type, or more generally, being unknown a priori. S^3-CIMA addresses this challenge by a weakly supervised learning model, i.e., a CNN model (Arvaniti and Claassen, 2017) that takes these sets of cell profiles as input and learns their association with the phenotype (**Fig. 1b**, see **Methods**).

Briefly, the model comprises a convolutional layer composed of filters, a pooling layer and a classification/regression output. The filters encode cell states by molecular profiles that are defined by fitting the model to the training data. The pooling layer computes cell frequency surrogates for the cell states encoded by each filter, which in turn are used to associate with the classification/regression phenotype.

The trained model is then used to identify the cell subsets critical for association with the phenotype (e.g., patient outcome, treatment response, survival). Therefore, the trained filter weights correspond to molecular profiles and not to predefined annotated cell types of cell subsets in the local anchor cell type neighborhood. These cell subsets are further characterized with respect to their cell type and molecular profile. The proximal enrichment of these cell subsets further motivates hypotheses about putative mechanisms conferring this enrichment (e.g., paracrine signaling).

We evaluated the S^3-CIMA workflow for studies based on two different *in situ* proteomic imaging technologies, i.e., CODEX highly multiplexed fluorescence microscopy data of 56 markers in tumor tissues of 35 patients with low- and high-risk CRC, as well as IMC data of 35 markers in pancreas tissues from 12 human donors with type 1 diabetes (T1D).

S^3-CIMA identifies spatially enriched cell subsets associated with differential CRC outcome

We applied S^3-CIMA to the data of the CRC study (Schürch et al., 2020) where patients were stratified in two groups with differential survival, i.e., 17 patients with Crohn's-like reaction (CLR) and 18 patients with diffuse inflammatory infiltration (DII). Overall survival of patients with CLR is significantly longer than that of patients with DII (Di Caro et al., 2014). All cells in the CODEX images had been annotated to 28 unique cell types including 18 immune, 6 stromal, 2 mixed and 1 tumor cell groups (Schürch et al., 2020). The study showed the

differences in the frequency of immune cell compositions between CLR and DII patients. Most notably, CLR patients had higher frequencies of B cells and lower frequencies of macrophages than DII patients, respectively (**Fig. S2**).

We expanded this analysis by assessing in an unbiased data-driven fashion the spatial enrichment of specific, possibly non-canonical cell subsets in the CRC TME between CLR and DII patients using the S^3-CIMA model (**Fig. 2a,b**).

We selected 100 random subsets of cells that are in k−NN (10 ≤ k ≤100) in each tissue image as the input for S^3-CIMA. Each input was labeled according to the image group as CLR or DII (**Fig. S3**). The classification model was trained on 12 samples from each CRC group and tested on the remaining 11 samples (**Fig. 2c**). The trained model was used to identify the *selected cells*, i.e., the cell subset whose frequency is relevant for classification of the CLR/DII patients (see **Methods**). The model robustly selected cells that are more frequent in the CLR group for neighborhood size k = 30 (p = 1.6e-12, Wilcoxon test) and in the DII group for neighborhood size k = 50 (p = 9.1e-13, Wilcoxon test) (**Fig. 2d**).

We assessed the cell type membership of the selected cells by calculating an enrichment score (ES) for the selected cells within their cell type compartment (see **Methods**). We found that at k = 30 the selected cells were predominantly from the CLR group and composed of tumor cells (16%) macrophages (14%), B cells (10%) and CD4+ T cells (9%) (**Fig. 2e**). The selected CD4+ T cells CD45RO+ and B cells were significantly enriched in the CLR group (**Fig. 2e**). The selected B cells comprise a B cell subset that exhibits higher expression of CD45RA compared to the non-selected cells (**Fig. 2f**). We identified that the frequencies of selected B cells, CD8+ T cells, plasma cells, stroma and vasculature are significantly higher in the CLR than DII group (**Fig. S3**), while the aforementioned cell type frequencies are not significantly different when all cells are considered (i.e., unsupervised analysis) (**Fig. S2**). On the other hand, we found that at k = 50 the selected cells were predominantly from the DII group and composed of macrophages (21%), tumor cells (17%), granulocytes (16%) and CD8+ T cells (8%) (**Fig. 2e**). These selected cells have higher expression of LAG3, PD-1 and VISTA compared to the bulk of their corresponding cell types that are in the tissue but not selected (**Fig. 2f**). The S^3-CIMA model can thus identify specific cell subsets at different neighborhood sizes which explain

different survival groups (i.e., at k=30 and k=50). This suggests that S^3-CIMA can capture the disease associated TMA characteristics at multiscale cellular neighborhoods. We also identified that the frequencies of selected tumor cells, CD4+ T cells CD45RO+, stroma and CD68+ macrophage are significantly higher in the DII than CLR group (**Fig. S3**).

S^3-CIMA reveals high enrichment of PD-1+ CD4+ T cell subset in granulocyte neighborhoods of DII patients

Schürch et al. (Schürch et al., 2020) defined nine cellular neighborhood classes as tissue regions with a high density of a specific cell type and assessed if the frequencies of these exhibit any differences between two CRC patient groups. This analysis showed that except for the follicle cellular neighborhood which was highly enriched in CLR patients, none of the other cellular neighborhood categories differed significantly. The study therefore suggested that to understand the underlying process shaping the CRC TME, cell types and the cellular neighborhood should be considered simultaneously.

S^3-CIMA allows to achieve this goal in an unbiased data-driven fashion, by learning disease associated cellular neighborhoods around specific anchor cell types, i.e., the *anchor cell niche* (**Fig. 1a**). We considered each cell type in the dataset as the anchor cell type. We found condition-specific cell type enrichments when using B cells, CD4+ T cells, tumor cells, M2 macrophages and granulocytes as anchor cell types across a range of k-NN sizes (10 ≤ k ≤50) (**Figs. S4-S9**).

Considering granulocytes as the anchor cell type, we found that the frequency of selected cells is significantly higher in the DII group compared to CLR and background at k = 30 (p = 4.6e-5 and p = 1.2e-15, Wilcoxon test, respectively) (**Fig. 3a**) and composed of granulocytes (58%), CD163+ macrophages (7%), tumor cells (6%) and smooth muscle (5%) (**Fig. 3b**).

To assess whether the selected cell types are spatially enriched in the granulocyte neighborhood and are not selected because of the abundance, we calculated an analytical ES for each cell type. The total number of cells of a given annotated cell type in the granulocyte neighborhood does not correlate with a higher ES (i.e., ES >1). The granulocyte niche is enriched by CD4+

T cells CD45RO+ (ES = 1.14), vasculature (ES = 1.05) and CD8+ T cells (ES = 1.04) (**Fig. 3b**). Interestingly, we confirmed that the granulocyte neighborhood is highly enriched for CD4+ T cells CD45RO+ in the DII group compared to the CLR group (**Fig. 3c**). To characterize the selected cell subset compared to non-selected cells and assess its differential functional capacity, we performed differential marker expression analysis. Specifically, we compared the functional marker expression, such as apoptosis, activation/proliferation, inhibition and cytokine signaling markers, in the selected subsets of cells versus non-selected cells within each specific cell type compartment (Fig. 3d, S10). The selected cells in the granulocytes niche highly express VISTA (KS = 0.48), Granzyme B (KS = 0.32), and MMP9 (KS = 0.31). We also identified that the spatially enriched CD4+ T cells CD45RO+ subset in the granulocyte niche over-expresses T cell exhaustion/activation markers VISTA (KS = 0.44), PD-1 (KS = 0.22) and EGFR (KS = 0.28) (**Fig. 3d**). To confirm the signaling activity of enriched cell subset in the granulocytes niche, we then mapped the selected cells in the granulocyte niche to the CODEX images and observed a high expression of PD-1 and CD4 markers in selected cell subsets (**Fig. 3e,f**).

Then, to investigate the mutual interaction of CD4+ T cells CD45RO+ and granulocytes, we performed S^3-CIMA local enrichment analysis considering CD4+ T cells CD45RO+ as the anchor cell type. We found that the frequency of selected cells is significantly higher in the CLR group compared to DII and background at k = 40 (p = 8.1e-8 and p = 0.019, Wilcoxon test, respectively) (**Fig. 3g**) and composed of mostly tumor cells (15%), CD163+ macrophages (15%) and B cells (8%) (**Fig. 3h**). The enrichment analysis showed that the CD4+ T cells CD45RO+ niche is enriched by smooth muscle (ES = 1.08), granulocytes (ES = 1.07), plasma cells (ES = 1.06) and CD8+ T cells (ES = 1.05) (**Fig. 3h**).

We then calculated the ES for each survival group separately and interestingly, we also identified the spatially enriched granulocyte subset in the CD4+ T cells CD45RO+ niche which is enriched in the DII group, indicating a specific mutual interaction of these two cell types (**Fig. 3i**). This interaction can be confirmed by colocalization of representative cells in the original CODEX images (**Fig. 3f**). These results suggest that S^3-CIMA can determine the local enrichment of the subset of a conventional cell type in the spatial neighborhood of a specific cell type as well as the one-to-one cell type interaction.

Functional spatial enrichment analysis reveals an EGFR-expressing macrophage subset in the proximity of granulocytes

We then performed *functional spatial enrichment analysis* by S^3-CIMA to identify cell subsets enriched in the proximity of granulocytes varying in local epidermal growth factor receptor (EGFR) expression. EGFR plays a key role in different cellular functions, such as proliferation, apoptosis, and differentiation. It has been shown that high expression of EGFR is common in many tumors. In particular in CRC, high expression of EGFR is associated with a poor prognosis (Cohen 2003).

To determine the effect of localized EGFR expression, we considered the phenotype of a multi-cell input (i.e., cells in the k nearest neighborhood) as the average expression of EGFR over all k cells in the nearest neighborhood of each granulocyte and then trained a regression S^3-CIMA model (R2-score = 63.8%, RMSE = 0.28) (**Fig. 4a**). To summarize the distribution of selected cells with respect to local EGFR expression, we report selected cell frequencies for either high/low, i.e., higher/lower than average expression across all considered cell neighborhoods. The model learns the spatial characteristics of both groups with two discriminative filters. We found that the frequency of selected cells is significantly higher in the $EGFR^{Low}$ group compared to $EGFR^{High}$ for filter 1 and vice versa for filter 2 at k = 20 (p = 2.2e-16, Wilcoxon test) (**Fig. 4b**). Considering the response of filter 1, we identified that the granulocyte niche is highly enriched in adipocytes, plasma cells and CD8+ T cells in the $EGFR^{Low}$ group (**Fig. 4c**). The spatially enriched CD8+ T cells subset in the granulocyte niche is associated with significantly reduced cytokine expression and proliferation activity (**Fig. 4d**) as it is known that activated granulocytes can inhibit cytokine production by T cells (Schmielau et. al., 2001). The response of filter 2 determined that the granulocyte niche is highly enriched by B cells and tumor cells in the $EGFR^{High}$ compared to the $EGFR^{Low}$ group (**Fig. 4c**). The spatially enriched B cells subset in the granulocyte niche significantly over-expresses the interleukin-2 receptor CD25 (KS = 0.35), EGFR (KS = 0.34) and PD-L1 (KS = 0.32) (**Fig. 4d**). This result showed that colocalization of granulocytes and B cells is associated with high expression of EGFR signaling in the TME, supporting the previous evidence for the role of granulocytes in the differentiation of neoplastic B cells (Costa et. al. 2019). With this case, we demonstrated that S^3-CIMA is capable of identifying determinants of niche composition that are associated with local signaling activity in the tissue such as EGFR signaling.

S^3-CIMA reveals high enrichment of beta cells in cytotoxic T cell neighborhoods in T1D onset patients

Type 1 diabetes is an autoimmune disease which is caused by immune system attack on insulin-producing beta cells in the pancreatic islets of Langerhans. The spatial interaction between islets and immune cells can be involved in the progression of the disease (Atkinson et al., 2014). We therefore examined if S^3-CIMA is capable of also finding a spatial pattern of cells in IMC data. To accomplish this, we used a type 1 diabetes (T1D) study including IMC measurement of 35 protein markers of pancreas tissues from 12 human donors, comprising healthy-, T1D onset- and long disease duration conditions (Damond et al. 2019) (**Fig. 5a**). This study demonstrated that beta cell destruction is preceded by recruitment of cytotoxic and helper T cells in T1D disease onset.

We performed S^3-CIMA analysis for anchor-based spatial enrichment analysis to understand association between islet cells and immune cells during T1D progression. The study showed that the association between immune cells and islet cells is not common; however, during T1D onset, because of the assembly of a destructive immune reaction, cytotoxic and T helper cells are recruited to beta cell-rich islets (Bluestone et al. 2020). Similarly, we also observed that the presence of immune cells (monocytes, neutrophils and B cells) in spatial proximity of islet cells (alpha, beta, gamma and delta cells) is not frequent in all three disease stages, i.e., there is no spatially enriched subset of islet cells when any of these immune cell types is an anchor cell. However, we observed that when we performed S^3-CIMA with T cells as an anchor (T helper cells, cytotoxic T cells and naïve T cells) (**Fig. S22**), there is a high enrichment of the presence of beta cells in the neighborhood of cytotoxic T cells from the T1D onset group (**Fig. 2k**). It is known that cytotoxic and helper T cells are involved in the destruction of beta cells in T1D (Boldison and Wong, 2016). This result confirms the previous finding of (Damond et al. 2019) that showed that beta cell destruction is preceded by recruitment of cytotoxic and helper T cells in T1D disease onset. In summary, we find that S^3-CIMA is capable of discovering novel disease-associated spatial cellular interactions from various spatial biology data types.

DISCUSSION

Understanding how cell type compositions in the TME change from one disease condition to another can pinpoint tentative disease mechanisms. S^3-CIMA provides a *supervised spatial enrichment analysis* by learning from disease-associated TME composition to systematically identify spatially enriched cell subsets associated with the disease conditions. We showed that S3-CIMA achieves this goal in an unbiased data-driven fashion, by learning disease associated cellular neighborhoods around specific anchor cell types (*local spatial enrichment analysis*) or random sets (*global spatial enrichment analysis*). We demonstrated, for both CODEX and IMC imaging data, that S^3-CIMA can reveal novel subpopulation structures and tissue organization possibly missed by single modalities or other methods. The frequency of selected subpopulations are significantly different across the phenotype groups (e.g., CLR and DII groups), while the frequency of the whole population of those cell types (i.e., unsupervised analysis) are similar in the phenotype groups (**Figs. S2-4**). Identified cellular subpopulations can, in principle, always be further subdivided to reflect a finer characterization of observed heterogeneity. Subdivision based on multiple modalities provides an opportunity to identify more meaningful biological distinctions.

Here, we examined the S^3-CIMA model on antibody-based proteomic imaging data that is inherently limited to subgenome-wide analyses. This limited spectrum of molecular parameters possibly precludes the detection of condition-associated cell subsets defined by unmeasured molecular markers. However, the increase in spatial resolution of spatial transcriptomic approaches is likely going to alleviate this limitation and allow S^3-CIMA to comprehensively assess the cell states across genome-wide expression profiles. Further, the greedy nature of the model fitting can fail to report the cell subsets that are less but still significantly correlated with the supervision signal than the selected cell subset.

Up to now, most of the spatial proteomic imaging data analysis efforts applied unsupervised methods that cannot use the associating external cues of interest directly. However, an increasing number of spatial proteomic imaging studies in the field of health and disease are and will be comparative, i.e., aiming at identifying spatial cell type composition patterns that explain external cues, such as physiological state, disease state, therapy response or signal transduction activity. We showed that S^3-CIMA can use both discrete and continuous states of these external cues as the sample phenotypes. S^3-CIMA constitutes a machine learning model

that addresses this issue and - as first of its kind - is expected to enable productive interpretation of such comparative studies in the future.

References

Arvaniti, E., Claassen, M., 2017. Sensitive detection of rare disease-associated cell subsets via representation learning. Nature Communications.

Baptista, A.P., Gola, A., Huang, Y., Milanez-Almeida, P., Torabi-Parizi, P., Urban Jr, J.F., Shapiro, V.S., Gerner, M.Y. and Germain, R.N., 2019. The chemoattractant receptor Ebi2 drives intranodal naive CD4+ T cell peripheralization to promote effective adaptive immunity. Immunity, 50(5), pp.1188-1201.

Black, S., Phillips, D., Hickey, J.W., Kennedy-Darling, J., Venkataraaman, V.G., Samusik, N., Goltsev, Y., Schürch, C.M. and Nolan, G.P., 2021. CODEX multiplexed tissue imaging with DNA-conjugated antibodies. Nature Protocols, 16(8), pp.3802-3835.

Bluestone, J. A., Herold, K., & Eisenbarth, G. 2010. Genetics, pathogenesis and clinical interventions in type 1 diabetes. Nature, 464(7293), 1293-1300.

Boldison, J. and Wong, F.S., 2016. Immune and pancreatic β cell interactions in type 1 diabetes. Trends in Endocrinology & Metabolism, 27(12), pp.856-867.

Cohen, R. B. (2003). Epidermal growth factor receptor as a therapeutic target in colorectal cancer. Clinical colorectal cancer, 2(4), 246-251.

Costa, S., Bevilacqua, D., Cassatella, M. A., & Scapini, P. (2019). Recent advances on the crosstalk between neutrophils and B or T lymphocytes. Immunology, 156(1), 23-32.

Damond, N., Engler, S., Zanotelli, V.R., Schapiro, D., Wasserfall, C.H., Kusmartseva, I., Nick, H.S., Thorel, F., Herrera, P.L., Atkinson, M.A. and Bodenmiller, B., 2019. A map of human type 1 diabetes progression by imaging mass cytometry. Cell metabolism, 29(3), pp.755-768.

Di Caro, G., Bergomas, F., Grizzi, F., Doni, A., Bianchi, P., Malesci, A., Laghi, L., Allavena, P., Mantovani, A. and Marchesi, F., 2014. Occurrence of tertiary lymphoid tissue is associated with T-cell infiltration and predicts better prognosis in early-stage colorectal cancers. Clinical Cancer Research, 20(8), pp.2147-2158.

Gerner, M.Y., Kastenmuller, W., Ifrim, I., Kabat, J. and Germain, R.N., 2012. Histo-cytometry: a method for highly multiplex quantitative tissue imaging analysis applied to dendritic cell subset microanatomy in lymph nodes. Immunity, 37(2), pp.364-376.

Giesen, C., Wang, H.A., Schapiro, D., Zivanovic, N., Jacobs, A., Hattendorf, B., Schüffler, P.J., Grolimund, D., Buhmann, J.M., Brandt, S. and Varga, Z., 2014. Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. Nature methods, 11(4), pp.417-422.

Goltsev, Y., Samusik, N., Kennedy-Darling, J., Bhate, S., Hale, M., Vazquez, G., ... & Nolan, G. P. 2018. Deep profiling of mouse splenic architecture with CODEX multiplexed imaging. Cell, 174(4), 968-981.

LeCun, Y., Bengio, Y. and Hinton, G., 2015. Deep learning. nature, 521(7553), pp.436-444.

Moses, L. and Pachter, L., 2022. Museum of spatial transcriptomics. Nature Methods, pp.1-13.

Palla, G., Spitzer, H., Klein, M., Fischer, D., Schaar, A.C., Kuemmerle, L.B., Rybakov, S., Ibarra, I.L., Holmberg, O., Virshup, I. and Lotfollahi, M., 2022. Squidpy: a scalable framework for spatial omics analysis. Nature methods, 19(2), pp.171-178.

Qi, H., Kastenmüller, W. and Germain, R.N., 2014. Spatiotemporal basis of innate and adaptive immunity in secondary lymphoid tissue. Annual review of cell and developmental biology, 30, pp.141-167.

Schmielau, J., & Finn, O. J. 2001. Activated granulocytes and granulocyte-derived hydrogen peroxide are the underlying mechanism of suppression of t-cell function in advanced cancer patients. *Cancer research*, *61*(12), 4756-4760.

Schürch, C.M., Bhate, S.S., Barlow, G.L., Phillips, D.J., Noti, L., Zlobec, I., Chu, P., Black, S., Demeter, J., McIlwain, D.R., Kinoshita, S., Samusik, N., Goltsev, Y., Nolan, G.P., 2020. Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. Cell 182, 1341–1359.e19.

Thornton, C. A., Mulqueen, R. M., Torkenczy, K. A., Nishida, A., Lowenstein, E. G., Fields, A. J., ... & Adey, A. C. 202. Spatially mapped single-cell chromatin accessibility. Nature communications, 12(1), 1-16.

Gizzi, A. M. C. 2021. A Shift in Paradigms: Spatial Genomics Approaches to Reveal Single-Cell Principles of Genome Organization. Frontiers in genetics, 12.

Lin, J.-R., Fallahi-Sichani, M., Chen, J.-Y. & Sorger, P. K. 2016. Cyclic immunofluorescence (CycIF), a highly multiplexed method for single-cell imaging. Curr. Protoc. Chem. Biol. 8, 251–264.

Salmén, F. et al., 2018. Barcoded solid-phase RNA capture for spatial transcriptomics profiling in mammalian tissue sections. Nat. Protoc. 13, 2501–2534.

Liu, Y. et al., 2020. High-spatial-resolution multi-omics sequencing via deterministic barcoding in tissue. Cell 183, 1665–1681.

Vickovic, S. et al. 2019. High-definition spatial transcriptomics for in situ tissue profiling. Nat. Methods 16, 987–990.

Dries, R., Zhu, Q., Dong, R., Eng, C. H. L., Li, H., Liu, K., ... & Yuan, G. C. 2021. Giotto: a toolbox for integrative analysis and visualization of spatial expression data. Genome biology, 22(1), 1-31.

METHODS

Notation used throughout the paper

C	Multi-cell input
(X,Y)	Position of a cell
k	Number of nearest neighbor cells
T	Cell type
p	p-value
k-NN	k-nearest neighbors
KS	Kolmogorov–Smirnov test statistic

Acronyms used throughout the paper

CRC	colorectal cancer
CLR	Crohn's-like reaction
DII	diffuse inflammatory infiltration
TME	tissue microenvironment
T1D	type 1 diabetes
ES	enrichment score
CNN	convolutional neural network
BG	back ground
MAD	median absolute deviation
EGFR	epidermal growth factor receptor

Datasets

Colorectal cancer

We used the CRC cohort including 56-marker multiplexed CODEX tissue imaging data of 35 CRC patients including 140 TMA spots i.e., 4 CODEX images for each patient. (Schürch et al., 2020). The flat table of the images was used for computing the spatial input of S^3-CIMA indicating 258,385 cells across 140 CODEX images of 35 patients as rows and patient id, marker expression, position (i.e., X/Y coordinates), survival group (i.e., DII or CLR) and annotated cell type (29 cell types) as columns. Marker intensity values were log transformed by adding an offset (log(1e-3+ x)) and scaled by z-score across the dataset prior to calculating spatial inputs. The spatial inputs of 24 patients (12 patients from each survival group) were randomly split into training and validation sets (80% and 20% respectively). The spatial input of the remaining 11 patients were used as the test. We calculated the nearest neighbors and enrichment analysis per TMA spots (i.e., images).

Type 1 diabetes

The type 1 diabetes (T1D) cohort includes IMC measurement of 35 protein markers of pancreatic islets of Langerhans tissues from 12 human donors, comprising healthy (n=4), recent-onset T1D (<0.5 years, n = 4) and long-standing T1D duration (>8 years, n = 4) (Damond et al. 2019). The data includes two sections originating from different anatomical regions of the pancreas (tail, body, or head) for each donor. The flat table of the IMC images was used for computing the spatial input of S^3-CIMA indicating 1,776,974 cells across 845 IMC images of 12 patients as rows and patient id, slide id (24 slides), cell position in each image (i.e., X/Y coordinates) cell category (exocrine, immune, islet, other, unknown), cell type (16 cell type) as columns. Marker intensity values were arcsinh transformed and scaled by z-score across the dataset prior to calculate spatial inputs. The spatial inputs of 9 patients (3 patients from each disease condition group) were randomly split into training and validation sets (80% and 20% respectively). The spatial input of the remaining 3 patients were used as the test set.

S^3-CIMA model training

S^3-CIMA implements a weakly supervised CNN model to identify cell subsets whose frequency distinguishes the considered phenotypes. The model is adopted from the CellCNN

model (Arvaniti and Claassen, 2017), comprising a single layer CNN, a pooling layer and a classification or regression output, and using groups of cell expression profiles (multi-cell inputs) as input. S3-CIMA uses multi-cell inputs, which are generated k nearest cell neighbors in a tissue image per phenotype with a set of marker expressions as the features (*global spatial enrichment analysis)*. For *anchor based spatial enrichment analysis,* S3-CIMA uses multi-cell inputs as k nearest cell neighbors around anchor cells of a specific cell type. Training and testing of the model are performed as described in (Arvaniti and Claassen, 2017). Briefly, 200 models were trained and the model with highest predictive accuracy on the validation set was selected.

Global spatial enrichment analysis

Global spatial enrichment analysis aims at identifying cell subsets that are spatially co-localized across the multiple conditions. The multi-cell input is generated using nearest neighbor cells of N randomly selected cells in each tissue image with a set of marker expressions as the features. A multi-cell input is the set of marker expression vectors of the k nearest neighbor cells $[\mathbf{C}_i]_{KM}$, $i \in \{1, 2, \cdots N\}$, where K and M are the number of cells and number of markers, respectively.

We performed S^3-CIMA global spatial enrichment analysis of the CRC cohort by selecting N random cells in each patient CODEX image and calculating k-NN $(10 \leq k \leq 100)$ based on the Euclidean distance. Each $\mathbf{C}_i$ input was labeled according to its corresponding image group as CLR or DII. Since some TMA images include only 200 cells, we selected N = 100 random cell subset as the spatial multi-cell inputs of the classification model.

Anchor based spatial enrichment analysis

Anchor based spatial enrichment aims at identifying cell subsets that are enriched in spatial proximity of a specific anchor cell type across the multiple conditions. The multi-cell inputs k nearest neighbors of an anchor cell are denoted $[\mathbf{C}_i]_{KM}$, $i \in \{1, 2, \cdots N\}$, where N, K and M are the number of the anchor cell, number of neighboring cells and number of markers, respectively. We added a set of randomly selected k nearest neighbor cells as the background (BG) phenotype to ensure that the cell signatures identified by the S^3-CIMA indicate the anchor cell specific spatial colocalization.

Functional spatial enrichment analysis

Functional spatial enrichment analysis aims at identifying cell subsets enriched in the proximity of a specific functional activity of a specific anchor cell type, e.g., local proliferative activity of tumor cells. We examined the S3-CIMA ability to detect spatial enriched functional subset using the colorectal cancer dataset. The preprocessing and computing of the multi-cell input $[C_i]_{KM}$ is similar to the anchor based spatial enrichment analysis. However, the multi-cell inputs were labeled according to the specific functional activity i.e., the average expression of a specific marker over all cells in $[C_i]_{KM}$. Here, the k-NN of that anchor cell type were computed, and the M-1 marker intensities (here, M = 49) of those selected k-cells were the features. The average of the functional marker intensities of the multi-cell inputs were used to capture the functional activity as the dependent variable during training of the model.

Quality of prediction of the regression model

In order to measure the accuracy and quality of the training, we used the score and RMSE in order to measure the predictive ability. The score, shows the amount of variance of the true values, which can be explained by the independent variable, and therefore gives a probability about how well unseen samples can be predicted. is defined as:

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n} (y, \hat{y})^2}{\sum_{i=1}^{n} (y, \underline{y})^2}$$

Where y is the true value and $\hat{y}$ is the predicted value of the sample from a total of n samples. The term $\underline{y}$ is the average of the true values.

The measure of RMSE, gives an idea of the standard deviation of the predictive error, which is the difference between the observed and predicted values. RMSE captures how spread out these values are. RMSE is obtained by taking the square-root of the mean squared error (MSE), another common metric for measuring the accuracy of the model.

Characterization of spatially enriched cell subsets

The downstream analysis using the trained filter weights facilitates the model interpretation and identifies subset of cells that are associated to the specific phenotype. The subset of cells with a high filter response indicates cells with a distinct signature to manifest the phenotype associated aspects of the TME. We select and perform downstream analyses for the cells with the positive filter response in most of the analysis unless stated. The relative frequency of

selected cells of all input cells per patient in each tissue type as well as background were calculated and compared using a Wilcoxon rank-sum test. The best neighborhood size (k) of each anchor was selected when achieving by considering two parameters i) the highest classification performance (validation and test accuracy), ii) the highest frequency of selected cell significance.

We assessed the cell type composition and enrichment for selected cell subsets in anchor based spatial enrichment analysis. Therefore, we defined an enrichment score that quantifies the association or exclusion of selected cells of specific cell type in the spatial proximity of the anchor cell.

We consider two quantities for calculating the enrichment score per cell type, i) frequency of cell type with high filter response in the spatial neighborhood of the anchor cells and ii) frequency of cell type with high filter response outside of the spatial neighborhood of the anchor cells:

We define the score of cell type T to be in the nearest neighborhood of the anchor cell and selected:

$$S_T^S = \frac{K_T^S}{E_T^S}$$

We define the score of cell type T to be in nearest neighborhood of the anchor cell:

$$S_T = \frac{K_T}{E_T}$$

where K_T^S is the number of selected cell type T in the nearest neighborhood of the anchor cell, K_T is the number of cell type T in the nearest neighborhood of the anchor cell, E_T and E_T^S are expected value of cell type T be presented in the nearest neighborhood of the anchor cell and expected value of cell type T be selected in the nearest neighborhood of the anchor cell, respectively and calculated as follow:

$$E_T^S = K^S \times \frac{N_T^S}{N^S}$$

$$E_T = K \times \frac{N_T}{N}$$

Where the the variables K being the number of all cells in the nearest neighborhood of the anchor cell, N the number of all cells in the image, N_T the number of all cells of cell type T in the image, K^S the number of all selected cells in the nearest neighborhood of the anchor cell, N^S the number of all selected cells and N_T^S the number of all selected cells of type T.

The enrichment score for cell type T is then defined by:

$$ES_T = \frac{S_T^S}{S_T}$$

The ES values above one indicates enrichment of the selected cells of type CT in the spatial neighborhood of the anchor cell.

The enrichment score of the global spatial enrichment analysis is calculated by:

$$E_T^S = N^S \times \frac{N_T}{N}$$

$$ES_T = \frac{N_T^S}{E_T^S}$$

The enrichment score of a specific cell type across the patient cohort is reported as the median value of the scores across all patients. The error bars are calculated as the median absolute deviation (MAD).

Differentially marker expression analysis

To characterize the spatially enriched cell subset, we examine if the functional marker's expression in selected subsets significantly differed from non-selected one. We quantify the difference in marker expression distribution by calculating the Kolmogorov–Smirnov two-sample test statistic (KS score) for each marker per cell type. The KS score equals to 1 indicates the highest effect size between two distributions and the KS score equals to 0 indicates that two distributions are similar.

Code availability

The code of the S^3-CIMA workflow has been deposited at https://github.com/claassenlab.

Figure captions

Fig. 1. S^3-CIMA overview and CRC global spatial enrichment analysis, S^3-CIMA leverages a single-layer CNN architecture adopted from the CellCNN model (Arvaniti and Claassen, 2017) **a.** The multi-cell input consists of k nearest neighbor cells around an anchor cell in a tissue image per phenotype and a set of marker expressions as the features. **b**. The model is a single-layer CNN in which node activities are computed by weighted sums over

marker expression values of each cell. The pooling layer summarizes the filter response of all cells in a multi-cell input per each convolutional filter. The trained filter weights are obtained by optimizing the CNN weights using true phenotypes of multi-cell inputs. For each cell in the image data, the filter response is calculated by scalar product of the cell marker expression and the CNN trained filter weights. The cells with relatively high filter response indicate cells with a distinct signature to manifest the phenotype of the image. The significant difference of frequency of selected cell populations between phenotypes shows that the CNN model learns the difference between the phenotypes. **c.** Different types of supervised spatial enrichment analysis including local spatial enrichment analysis for identifying which cell subsets are enriched in the proximity of an anchor cell type, functional spatial enrichment analysis for identifying cell subsets enriched in the proximity of a specific functional activity of the tissue. **d**. Two types of TME definitions: unsupervised and supervised. unsupervised approaches allow for defining cell type composition of the TME, supervised approaches go a step further to identify differential cell type composition across tissue conditions, including condition associated cell subtypes.

Fig. 2. S^3-CIMA global spatial enrichment analysis a. S^3-CIMA global spatial enrichment analysis applied on the 140 CODEX images of 35 CRC patients. **b**. Uniform Manifold Approximation and Projection (UMAP), indicating different cell types in the CRC tumor microenvironment. UMAP plots showing origin of single cells by clinical group of patients (CLR, DII) and expression of indicated markers. **c.** S^3-CIMA classification model performance (test and train accuracy) and -log(p-value) of comparing the frequency of selected cells by model across different cell neighborhood sizes (10 to 100). **d**. The frequency of selected cells between groups in k = 30 (significantly more from CLR group) and k = 50 (significantly more from DII group) are shown. Selected cells (colored by cell type, color legend Fig.2b) are mapped back to the corresponding patient CODEX images in both CLR and DII groups (also Fig. S21). **e.** The frequency and enrichment score of selected cells across the cell types. Bars are colored by cell type. **f.** Density of the marker expression showing greatest differential abundance in terms of the Kolmogorov–Smirnov two-sample test statistics between the selected and non-selected cell subsets in all cell populations and per cell types. The high differential abundance (KS statistics) is highlighted by yellow color.

Fig. 3. S^3-CIMA local (anchor based) spatial enrichment analysis, The multi-cell inputs generated from k nearest neighbor cells around granulocytes (k = 30) and CD4+ T cells

CD45RO+ (k = 40) across all 35 CODEX images of the CRC data set. Each multi-cell input was labeled by a CRC outcome class (CLR or DII) for the classification task. **a.** Frequency of selected cells across the patients between groups with granulocytes as an anchor. The cells with the high filter response have significantly increased frequency in the DII group **b.** Frequency and enrichment score (ES) of selected cells across the cell types obtained from local enrichment spatial analysis for granulocytes as the anchor. Bars are colored by cell type. **c.** Bubble plot of enrichment score of each cell type in the proximity of granulocytes comparing different survival groups (median of enrichment scores across all patients). The MAD is shown by the error bar. The color and size of the bubbles indicate the cell types (color legend Fig.2b) and the ratio of the number of selected cells to the total number of cells of the specific type, respectively. **d**. Selected cells (colored by cell type, color legend Fig.2b) are mapped back to the corresponding patient CODEX images in the DII groups (Fig. S19). **e.** Density of functional marker expression showing greatest differential abundance in terms of the Kolmogorov–Smirnov two-sample test between the selected and non-selected cell subsets in all cell populations and per cell types for classification task. **f.** Codex image of patient 30 (DII group). The spatial neighborhood of one granulocyte as the anchor is zoomed in. The PD-1, CD4, CD15 and CD68 markers intensities of the selected cells around this anchor are shown. **g.** Frequency of selected cells across the patients between groups with CD4+ T cells CD45RO+ as an anchor. The cells with the high filter response have significantly increased frequency in the CLR group. **h.** Frequency and enrichment score (ES) of selected cells across the cell types obtained from local enrichment spatial analysis for CD4+ T cells CD45RO+ as the anchor. Bars are colored by cell type. **i.** Bubble plot of enrichment score of each cell type in the proximity of CD4+ T cells CD45RO+ comparing different survival groups (median of enrichment scores across all patients). The MAD is shown by the error bar. The color and size of the bubbles indicate the cell types (color legend Fig.2b) and the ratio of the number of selected cells to the total number of cells of the specific type, respectively.

Fig. 4. S^3-CIMA functional spatial enrichment analysis, a. The multi-cell inputs generated from k nearest neighbor cells around granulocytes (k = 20) across all 35 CODEX images of the CRC data set. Each multi-cell input was labeled by the average of EGFR marker expression of the cells in that multi-cell input (continuous value) as a surrogate for local signal transduction activity for the regression task. **b.** Frequency of selected cells of functional enrichment analysis with granulocytes as an anchor comparing two groups of patients with high and low EGFR marker expression. The cells with the high filter response have significantly increased

frequency in the $EGFR^{Low}$ group in filter 1 and $EGFR^{High}$ group in filter 2. **c.** Bubble plot of enrichment score of each cell type in the proximity of granulocytes showing the cytokine signaling activity of the EGFR marker in each group. The MAD is shown by the error bar. The color and size of the bubbles indicate the cell types (color legend Fig.2b) and the ratio of the number of selected cells to the total number of cells of the specific type, respectively. **d**. Selected cells (colored by cell type, color legend Fig.2b) and the EGFR expression value for each cell are mapped back to the corresponding patient CODEX images in both $EGFR^{High}$ and $EGFR^{Low}$ groups. **e.** Density of functional marker expression showing greatest differential abundance in terms of the Kolmogorov–Smirnov two-sample test between the selected and non-selected cell subsets in all cell populations and per cell types for regression task.**f.** Codex images of Patient 16 (Spot 32_A, DII group, low EGFR) and Patient 17 (Spot 33_B, CLR group, high EGFR). The spatial neighborhood of one granulocyte as the anchor is zoomed in. The mentioned markers intensities of the selected cells around this anchor are shown.

Fig. 5. S^3-CIMA local (anchor based) spatial enrichment analysis a. S^3-CIMA anchor based spatial enrichment analysis was applied on IMC measurement of 35 protein markers of pancreas tissues from 12 human donors, comprising healthy-, T1D onset- and long disease duration conditions. **b.** Frequency of selected cells between disease stages with helper T cells as an anchor. The cells with the high filter response have significantly increased frequency in the onset T1D group. Frequency of selected cells between disease stages with cytotoxic T cells as an anchor. The cells with the high filter response have significantly increased frequency in the non-diabetic group. **c.** Bubble plot of enrichment score of each cell type in the proximity of helper T cells shows that there is no high enrichment of the presence of any cell type. The color and size of the bubbles indicate the cell types and the ratio of the number of selected cells to the total number of cells of the specific type, respectively. Bubble plot of enrichment score of each cell type in the proximity of cytotoxic T cells shows that there is a high enrichment of the presence of beta cells from the onset T1D group. **d.** Density of functional marker expression showing greatest differential abundance in terms of the Kolmogorov–Smirnov two-sample test between the selected and non-selected cell subsets in all cell populations and beta cells.

Fig. 1

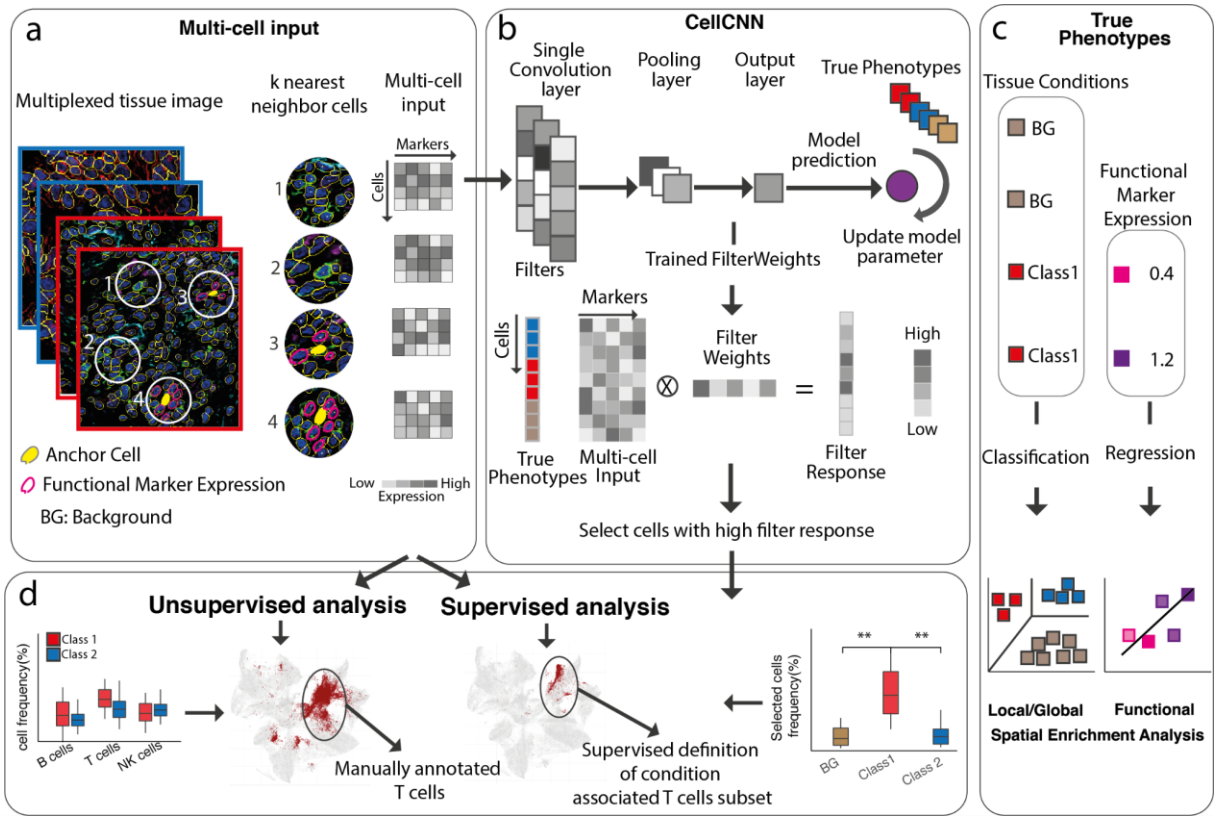

a
Multi-cell input
Multiplexed tissue image
k nearest neighbor cells
Multi-cell input
Markers
Cells
Anchor Cell
Functional Marker Expression
BG: Background
Low
High
Expression
b
CellCNN
Single Convolution layer
Pooling layer
Output layer
True Phenotypes
Model prediction
Update model parameter
Filters
Trained FilterWeights
Markers
Cells
Filter Weights
High
Low
True Phenotypes
Multi-cell Input
Filter Response
Select cells with high filter response
c
True Phenotypes
Tissue Conditions
BG
BG
Class1
Class1
Functional Marker Expression
0.4
1.2
Classification
Regression
Local/Global
Functional
Spatial Enrichment Analysis
d
Unsupervised analysis
Supervised analysis
Class 1
Class 2
cell frequency(%)
B cells
T cells
NK cells
Manually annotated T cells
Supervised definition of condition associated T cells subset
Selected cells frequency(%)
**
**
BG
Class1
Class 2

Fig. 2

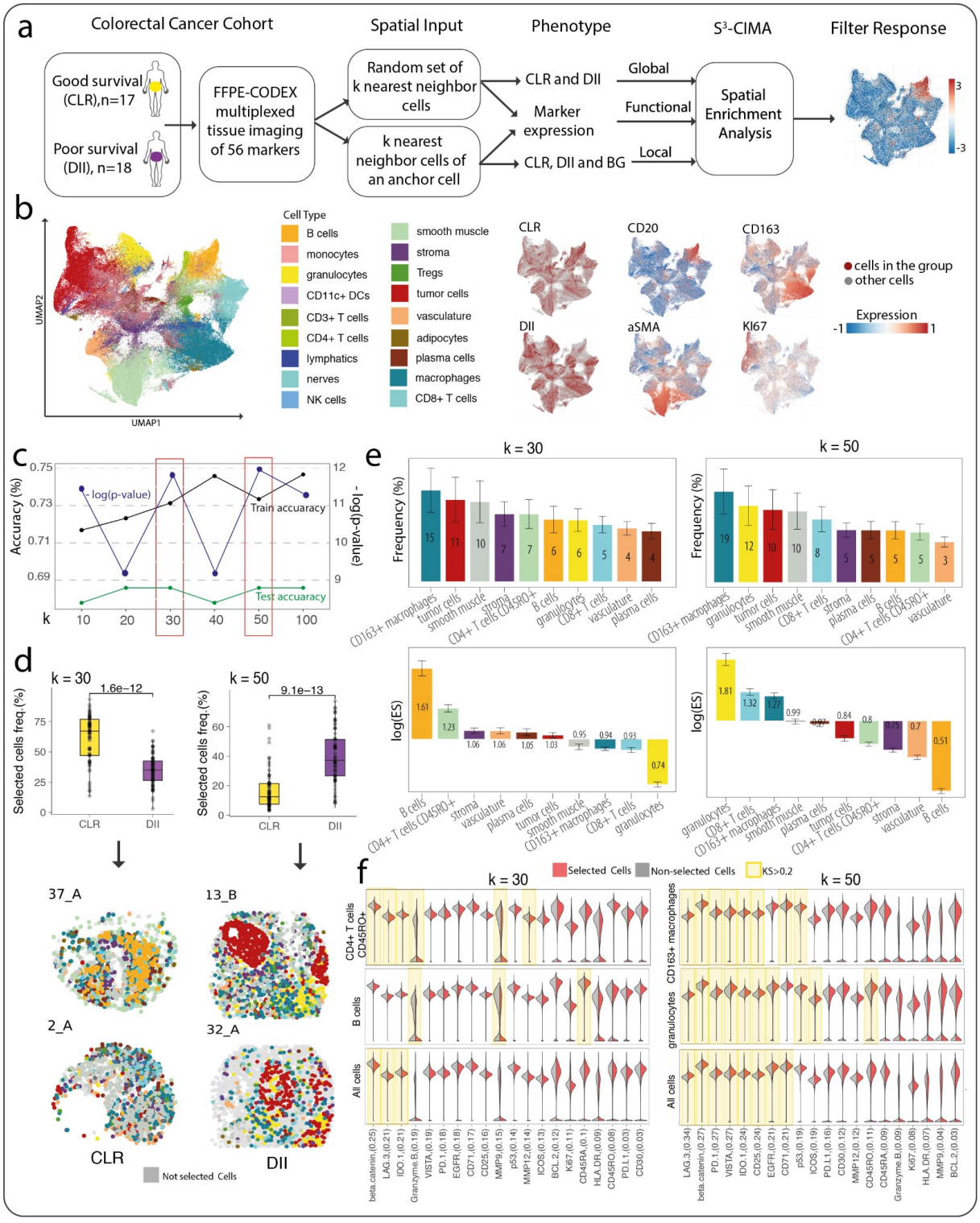

a
Colorectal Cancer Cohort
Spatial Input
Phenotype
S³-CIMA
Filter Response
Good survival (CLR),n=17
Poor survival (DII), n=18
FFPE-CODEX multiplexed tissue imaging of 56 markers
Random set of k nearest neighbor cells
k nearest neighbor cells of an anchor cell
CLR and DII
Marker expression
CLR, DII and BG
Global
Functional
Local
Spatial Enrichment Analysis
b
UMAP2
UMAP1
Cell Type
B cells
monocytes
granulocytes
CD11c+ DCs
CD3+ T cells
CD4+ T cells
lymphatics
nerves
NK cells
smooth muscle
stroma
Tregs
tumor cells
vasculature
adipocytes
plasma cells
macrophages
CD8+ T cells
CLR
CD20
CD163
DII
aSMA
KI67
cells in the group
other cells
Expression
c
Accuracy (%)
log(p-value)
Train accuracy
Test accuracy
-log(p-value)
k
e
k = 30
k = 50
Frequency (%)
log(ES)
d
k = 30
k = 50
1.6e−12
9.1e−13
Selected cells freq.(%)
CLR
DII
37_A
13_B
2_A
32_A
Not selected Cells
f
Selected Cells
Non-selected Cells
KS>0.2
CD4+ T cells CD45RO+
B cells
All cells
CD163+ macrophages
granulocytes

Fig.3

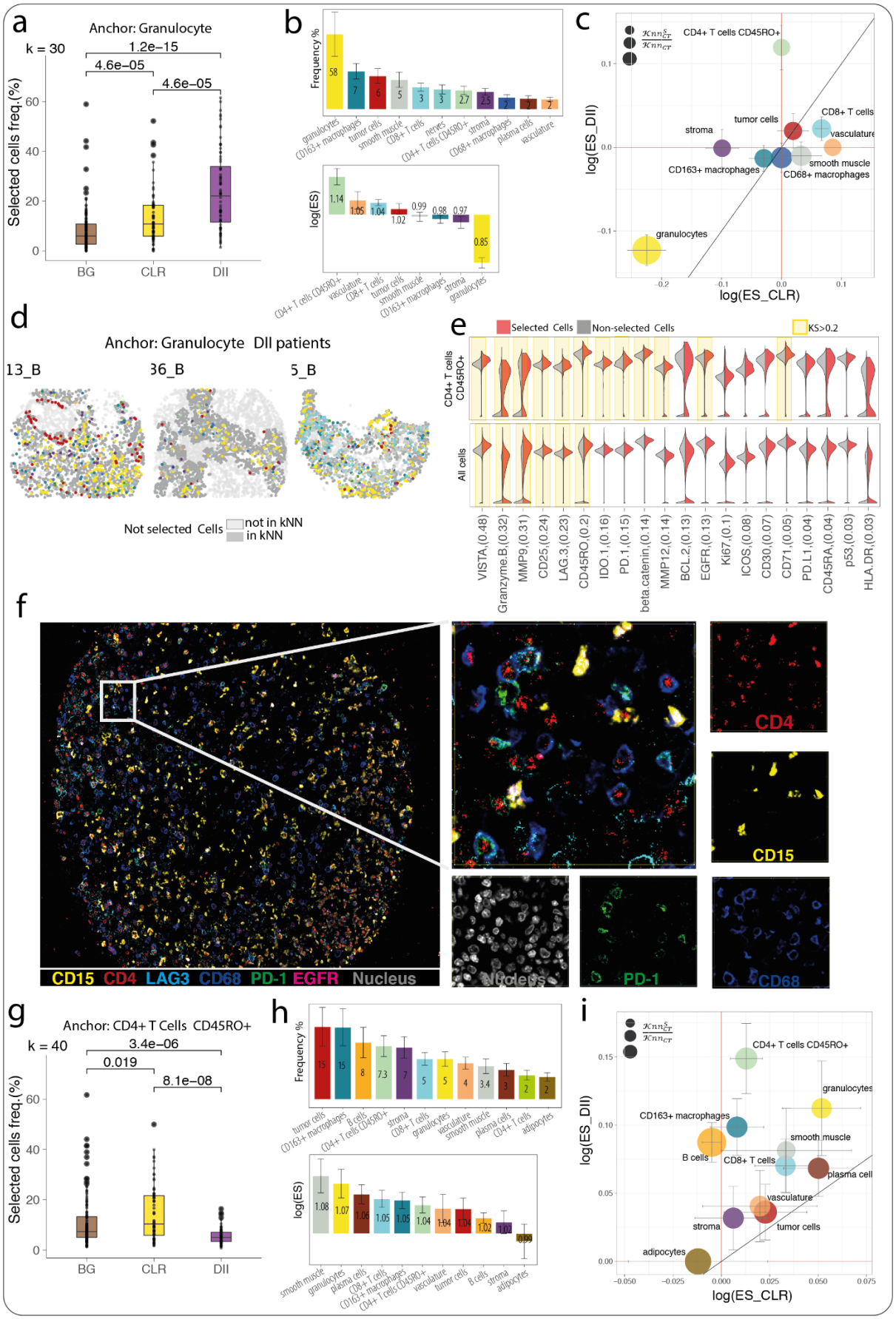
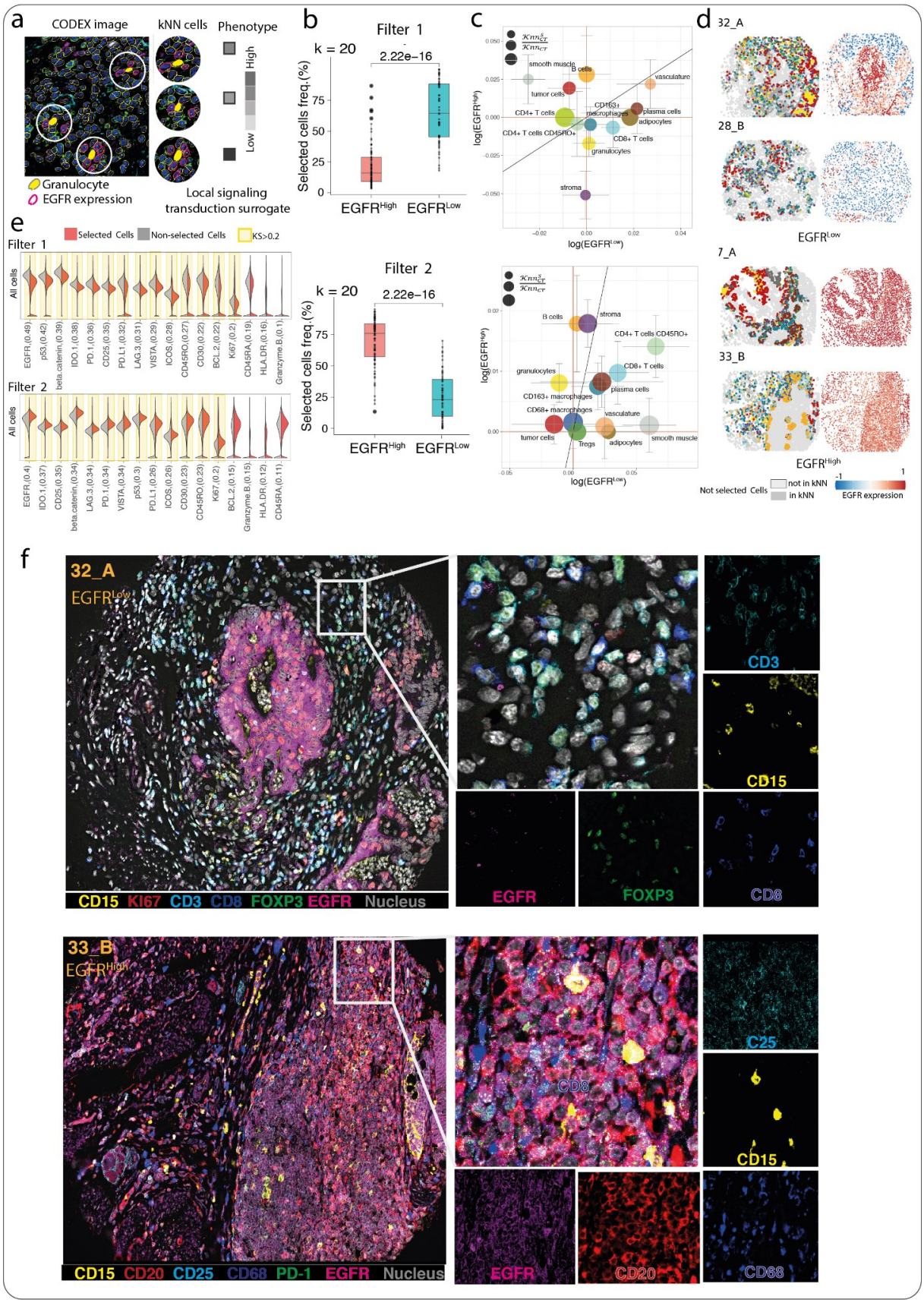

a
Anchor: Granulocyte
k = 30
1.2e−15
4.6e−05
4.6e−05
Selected cells freq.(%)
BG
CLR
DII
b
Frequency %
log(ES)
c
log(ES_DII)
log(ES_CLR)
CD4+ T cells CD45RO+
tumor cells
stroma
CD8+ T cells
vasculature
CD163+ macrophages
smooth muscle
CD68+ macrophages
granulocytes
d
Anchor: Granulocyte DII patients
13_B
36_B
5_B
Not selected Cells
not in kNN
in kNN
e
Selected Cells
Non-selected Cells
KS>0.2
CD4+ T cells CD45RO+
All cells
VISTA,(0.48)
Granzyme.B,(0.32)
MMP9,(0.31)
CD25,(0.24)
LAG.3,(0.23)
CD45RO,(0.2)
IDO.1,(0.16)
PD.1,(0.15)
beta.catenin,(0.14)
MMP12,(0.14)
BCL.2,(0.13)
EGFR,(0.13)
Ki67,(0.1)
ICOS,(0.08)
CD30,(0.07)
CD71,(0.05)
PD.L1,(0.04)
CD45RA,(0.04)
p53,(0.03)
HLA.DR,(0.03)
f
CD15 CD4 LAG3 CD68 PD-1 EGFR Nucleus
CD4
CD15
Nucleus
PD-1
CD68
g
Anchor: CD4+ T Cells CD45RO+
k = 40
3.4e−06
0.019
8.1e−08
Selected cells freq.(%)
BG
CLR
DII
h
Frequency %
log(ES)
i
log(ES_DII)
log(ES_CLR)
CD4+ T cells CD45RO+
granulocytes
CD163+ macrophages
smooth muscle
B cells
CD8+ T cells
plasma cell
vasculature
stroma
tumor cells
adipocytes

Fig. 4


a
CODEX image
kNN cells
Phenotype
High
Low
Granulocyte
EGFR expression
Local signaling
transduction surrogate
b
Filter 1
k = 20
2.22e−16
Selected cells freq.(%)
EGFR^High
EGFR^Low
Filter 2
k = 20
2.22e−16
c
smooth muscle
B cells
vasculature
tumor cells
CD163+ macrophages
plasma cells
CD4+ T cells
adipocytes
CD4+ T cells CD45RO+
CD8+ T cells
granulocytes
stroma
log(EGFR^High)
log(EGFR^Low)
d
32_A
28_B
EGFR^Low
7_A
33_B
EGFR^High
Not selected Cells
not in kNN
in kNN
EGFR expression
e
Selected Cells
Non-selected Cells
KS>0.2
Filter 1
Filter 2
All cells
EGFR,(0.49)
p53,(0.42)
beta.catenin,(0.38)
IDO.1,(0.38)
PD.1,(0.36)
CD25,(0.35)
PD.L1,(0.32)
LAG.3,(0.31)
VISTA,(0.29)
ICOS,(0.28)
CD45RO,(0.27)
CD30,(0.22)
BCL.2,(0.22)
Ki67,(0.2)
CD45RA,(0.19)
HLA.DR,(0.16)
Granzyme B,(0.1)
EGFR,(0.4)
IDO.1,(0.37)
CD25,(0.35)
beta.catenin,(0.34)
LAG.3,(0.34)
PD.1,(0.34)
VISTA,(0.34)
p53,(0.3)
PD.L1,(0.26)
ICOS,(0.26)
CD30,(0.23)
CD45RO,(0.23)
Ki67,(0.2)
BCL.2,(0.15)
Granzyme B,(0.15)
HLA.DR,(0.12)
CD45RA,(0.11)
f
32_A
EGFR^Low
CD15 KI67 CD3 CD8 FOXP3 EGFR Nucleus
CD3
CD15
EGFR
FOXP3
CD8
33_B
EGFR^High
CD15 CD20 CD25 CD68 PD-1 EGFR Nucleus
CD8
C25
CD15
EGFR
CD20
CD68