Department of Mathematics
Chair of Mathematical Modeling of Biological Systems

Predicting transcription rate from multiplexed protein maps using deep learning

Master's Thesis by Andres Becker

Examiner:	Prof. Dr. Fabian J. Theis
Advisor:	Dr. Hannah Spitzer
Submission Date:	May 14, 2021

Declaration

I hereby declare that this thesis is my own work and that no other sources have been used except those clearly indicated and referenced.

Andres Alberto Becker Sanabria, München, 14.05.2021

Abstract

By means of fluorescent antibodies it is possible to observe the amount of nascent RNA within the nucleus of a cell, and thus estimate its Transcription Rate (TR). But what about the other molecules, proteins, organelles, etc. within the nucleus of the cell? Is it possible to estimate the TR using only the shape and distribution of these subnuclear components? By means of multichannel images of single cell nucleus (obtained through the multiplexed protein map (MPM) protocol [GHP18]) and Convolutional Neural Networks (CNNs), we show that this is possible. Applying pre-processing and data augmentation techniques, we reduce the information contained in the intensity of the pixels and the correlation of these between the different channels. This allowed the CNN to focus mainly on the information provided by the location, size and distribution of elements within the cell nucleus. For this task different architectures were tried, from a simple CNN (with only 160k parameters), to more complex architectures such as the ResNet50V2 or the Xception (with more than 20m parameters). Furthermore, through the interpretability methods Integrated Gradient (IG) and VarGrad (VG), we could obtain score maps that allowed us to observe the pixels that the CNN considered as relevant to predict the TR for each cell nucleus input image. The analysis of these score maps reveals how as the TR changes, the CNN focuses on different proteins and areas of the nucleus. This shows that interpretability methods can help us to understand how a CNN make its predictions and learn from it, which has the potential to provide guidance for new discoveries in the field of biology.

Acknowledgments

To my father, my partner in my wildest adventures, best friend and who taught me what are the important things in life. He may never read this, but let the world known he is a loved and admired man.

Contents

Chapter 1

Introduction

1.1 Motivation

Understanding how RNA concentration in eukaryotes cells is regulated is very important to understand gene expression. However, measuring the amount of RNA inside a cell, may not be enough to fully describe cellular function. Accordingly to Buxbaum et al. [BHS14] and Korolchuk et al. [Kor+11], cellular function can heavily depend on the specific intracellular location and interaction with other molecules and intracellular structures. According to Vogel et al. [Vog+10], explaining gene expression using only nuclear protein abundance can be done only until a certain extent.

In recent years the development of new methods for capturing spatial organization and distribution of several subnuclear bodies and proteins/molecules in high resolution images [GHP18], makes the use of Convolutional Neural Network (CNN) models to analyze this information a natural choice. As CNNs have proven to be powerful tools in the recognition of spatial patterns [KSH12]. The use of CNNs not only would allow us to estimate gene expression. The continuous development of interpretability methods [Hoo+18], would also allow us to understand how these models work and learn from them.

For these reasons, the use of models capable of focusing on subnuclear spatial information, can potentially help us to understand better cell expression.

1.2 Goal of this thesis

The main objective of this work is to prove that it is possible to predict the Transcription Rate (TR) of a cell, using a deep learning model and the spatial information of its nucleus. More specifically, this means to design a Convolutional Neural Network (CNN) architecture for regressing the TR value of a cell, using only the information encoded in the location, distribution and shape of subnuclear components (like molecules, proteins and nuclear bodies) in multichannel images of cell nucleus. In order to do it, we implement preprocessing and data augmentation techniques aimed to reduce the information contained in the intensity of the pixels and its correlation among the channels. This would encourage the model to focus mainly on spatial information.

The second goal of this work is to apply recent gradient-based interpretability methods [Ade+20], to understand which molecules/proteins and nuclear bodies were most relevant for the prediction of the mode. Understanding how a CNN model works gives us the possibility to learn from it, which has the potential to provide guidance for new discoveries in the field of biology.

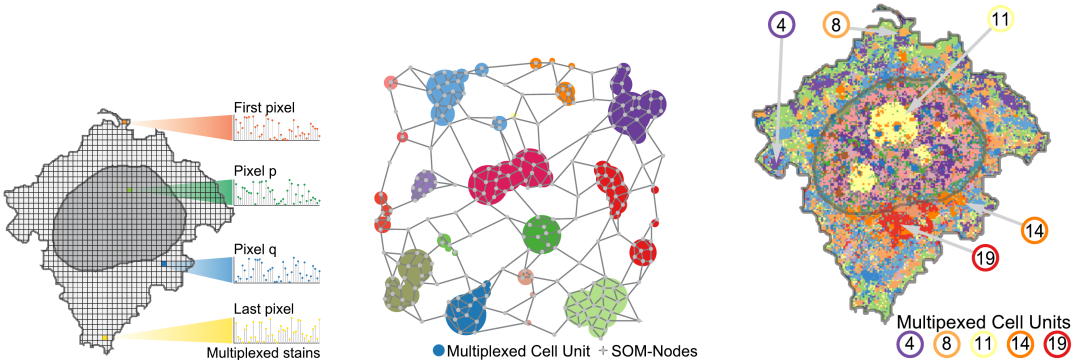
1.3 Literature review

Can we fully describe gene expression using only information about concentration of proteins and/or molecules like RNA inside the cell nucleus? Accordingly to Buxbaum et al. [BHS14], the location of messenger RNA (mRNA) within the cell plays an important role in protein synthesis. In [Kor+11], Korolchuk et al. show that cellular response to nutrient levels is a mechanism that needs to contemplate the position of Lysosomes (dynamic intracellular organelles) in order to be fully understood. However, the need for localization information to explain cellular mechanisms is not only limited to a subcellular level, but also at a subnuclear level. For instance, in [SF19] van Steensel et al. argue that the spatial organization of subnuclear components can have an important role in the regulation of gene expression. In [Vog+10], Vogel et al. shows that in human cells the concentration of mRNA can only explain protein abundance to a certain extent, which could indicate the need to consider spatial information to predict protein expression.

In recent years, the implementation of new imaging technologies has made it possible to access subnuclear spatial information. In [GHP18], Gut et al. introduce the iterative indirect immunofluorescence imaging (4i) protocol, which is a process that allows to efficiently capture thousands of single cell multichannel images from a cell culture without degrading it. The 4i protocol is part of the multiplexed protein map (MPM) protocol also introduced in [GHP18], which allows the segmentation of the tissue images into single cell nucleus images (Multiplexed single cell analysis). The MPM protocol also introduces two other features that are not used in this work, but are still worth mentioning. The first one is the multiplexed cell unit (MCU) analysis, which segments the cell nucleus image into regions. These regions can be then used to identify subnuclear bodies or protein complexes. The segmentation is done through two unsupervised clustering algorithms[1], applied over the measured pixel intensities. The MCU analysis is shown on figure 1.1.

The second feature of the MPM protocol that is not discussed here, but could be used in future work, is the application of pharmacological and metabolic perturbations to some sections of the cell culture. The analysis shown in [GHP18] revealed expected and unexpected changes in the concentration and distribution of molecules inside the

[1]To identify clusters in an unsupervised manner, *Self Organizing Maps* algorithm and *Phenograph* analysis are used over a very large number of pixels sampled from all the single cells images.



(a) Extraction of pixel intensities.

(b) Pixel clustering by Self Organizing Maps and Phenograph.

(c) Cell subdivision base on the MCU.

Figure 1.1: Figure a shows the pixel intensity extraction for a single cell. The pixel intensity is a vector containing the readout of that 2D location for each protein, one specific protein readout per entrance. Figure b shows the clusters found by Self Organizing Maps algorithm and Phenograph analysis over the pixel intensities. Figure c shows a cell masked with the clusters found by the MCU analysis. Images source [GHP18].

cell.

Artificial Neural Network (ANN) are very robust tools widely used in the field of Machine Learning (ML) that can potentially approximate any function [Cyb89], [HSW89], [Fun89]. In the field of biology, ANNs have proven capable of solving very complex and high-impact problems. One of the best examples in recent years is the three-dimensional prediction of the structure of a protein using amino acid sequences encoded in the genes [Sen+20], which is a very important problem since the structure of a protein largely determines its function. In [Che+16], Chen et al. introduced a deep ANN model known as *D-GEX*, which outperformed previous linear model approaches when trained using gene expression profiling data.

In [KSH12] Krizhevsky et al. show that Convolutional Neural Networks (CNNs) are powerful tools in the recognition of spatial patterns, achieving outstanding results in ImageNet LSVRC-2010 contest. This makes CNNs suitable models to analyze spatial information embedded in images of single cell nucleus, like the ones provided by the MPM protocol.

However, in many fields of study and industries, the interpretation of the models is essential. For example, in the medical field, CNN architectures have achieved remarkable results in the segmentation of brain tumors [SSR21]. However, to successfully implement deep learning models in the diagnosis of patients, it is not enough only to know what the model predicts, but also how it does it.

Many researchers have proposed different techniques to explain what happens inside black-box models like CNNs. The difference between these methods is basically whether they are applicable to any type of model (model-agnostic/model-independent) or only to a specific group (model-specific). An example of a model-independent method is the Local Interpretable Model-Agnostic Explanations (LIME), which basically aims to approximate the underlying model f (not interpretable) by means of an interpretable model g (e.g. a linear model) for a specific region of the input [RSG16]. As the name suggest, LIME provides a local and individual explanation of each input. However, there are other methods that provide a general (global) explanation of the model. An example of a global method (and also model-specific) would be the visualization of the learned filters/kernels of a CNN, which can indicate the features in the data that are important for the model prediction [ZF14].

However, in this work we use *attribution methods*, which are aimed to rank each input feature based on how much they contribute to the output of the model. These methods create an importance (or score) map for each element of the input data. There are several ways to compute these score maps [Bae+10], [Shr+16]. However, most of these methods base the importance assignment of each input feature on the gradient of the model output with respect to the input (gradient-based methods) [SVZ13], [Bin+16] and [Spr+14].

Nevertheless, just using the gradient as a feature importance designation method is not enough. As a model learns the relationship between an input and its output, the gradient of the model's output with respect to the input features will approximate to 0 (saturation). To alleviate this issue, Sundararajan et al. [STY17] proposed Integrated Gradient (IG), which accumulates the gradient of the output with respect to the input when it goes from an uninformative value to the original input.

However, in practice attribution methods like IG often produce noisy and diffuse score maps, and in some cases they are not even better than a random designation of feature importance [Hoo+18]. For this reason Smilkov et al. [Smi+17] proposed an ensemble interpretability method known as SmoothGrad (SG), which in practice reduces noise in score maps and can be easily combined with other attribution methods such as IG. In this work we use a slightly different version proposed by Adebayo et al. [Ade+18] known as VarGrad (VG), which is inspired by SG and has been shown to empirically outperform such a random assignment of importance [Hoo+18].

Chapter 2

Basics

This chapter provides a theoretical explanation of all the elements used in this work. It is divided into 3 main sections

1. The biological background
2. The Machine Learning basics
3. the Interpretability methods background

The first part provides a brief explanation of what cell expression is, as well as the transcription process, which is one of its main parts and the central subject of this work. The second part is a short explanation of what Artificial Neural Networks (ANNs) are, specifically Convolutional Neural Networks (CNNs) and their different components. It also provides a short explanation of the basic concepts needed to understand the idea behind the pre-built architectures used in this work, the ResNet50V2 and the Xception. Finally, the third part explains the methods used to interpret the results of the trained models. This interpretability methods are aimed to rank each input feature based on how much they contribute to the output of the model.

The theory behind the dataset used in this work is not provided in this chapter. Instead, a whole chapter was dedicated to it (see chapter 3). Chapter 3 also introduces the preprocessing and data augmentation techniques used in this work.

2.1 Biological Background

Cells are considered the smallest unit of life. There are two types of cells, *prokaryotic* and *eukaryotic*. The main difference between these, is that prokaryotic cells do not contain nucleus and that that prokaryotes are considered single-celled organisms, while eukaryotes organisms can be either single-celled or multicellular. For multicellular organisms, like plants or mammals, eukaryotic cells are the *building-blocks* of life. This work focuses on a process of eukaryotic cells. Therefore, in the subsequent when we refer to cells, we will be referring to eukaryotic cells only.

Multicellular organisms (like us) have different cell types, where each one of them can have many or an specific function. For instance, red blood cells are responsible for

carrying the oxygen in the body. In order to carry as much oxygen as possible they lack a nucleus, and therefore they are unable to undergo *mitosis*[1].

However, there are also cells aimed to produce (*synthesize*) certain substances that regulate process in our body. For instance, *Alpha cells* are pancreatic cells responsible for synthesizing the *glucagon* hormone, which elevates the glucose levels in the blood [Ker99]. The process in which cells produce this substances is called *cellular expression* or *gene expression*. The reason why this process is also called gene expression, is because the instructions to synthesize every substance (or any functional product, like hormones or proteins) are encoded in a specific gene[2].

There are two key steps involved in gene expression, *transcription* and *translation*. Roughly speaking, transcription is the process in which the instructions to synthesize a product (like proteins) are copied from a gene in the DNA, to a single strand molecule called messenger RNA (mRNA). On the other hand, Translation is the process in which the instructions in the mRNA are interpreted to produce a functional product. Figure 2.1 shows a simple representation of this process.

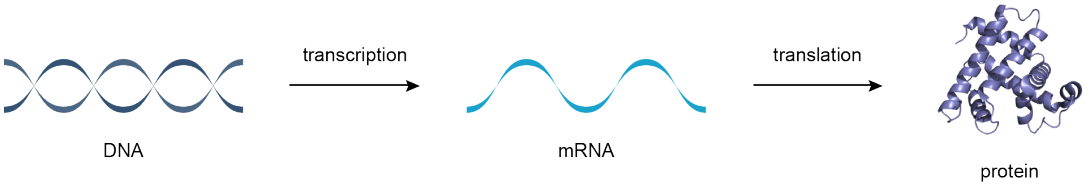


Figure 2.1: Simple representation of the gene expression process. Image source [BJ].

The transcription process happens inside the cell nucleus, while translation happens in the *ribosome* (outside the nucleus). The reason why transcription is necessary, is because the instructions needed to build a product are encoded in the DNA, which is inside the nucleus. Since DNA is too big to pass the membrane that covers the nucleus (nuclear envelop) to travel to the ribosome (which is the organelle in charge of building the product), the necessary instructions in the DNA are copied into a smaller strand (mRNA), which is now able to escape the nucleus and travels to the ribosome to start the translation process. Figure 2.2 shows a diagram of an eukaryotic cell and some of its parts. This work focuses on the transcription process and the factors that seed up or slow down this process.

[1] Mitosis is the process through which eukaryotic cells reproduce themselves and give rise to new organisms.

[2] A *gene* is defined as a region of the *DNA* that encodes a function. DNA is contained in *chromosomes*, which are long DNA strands containing many genes.

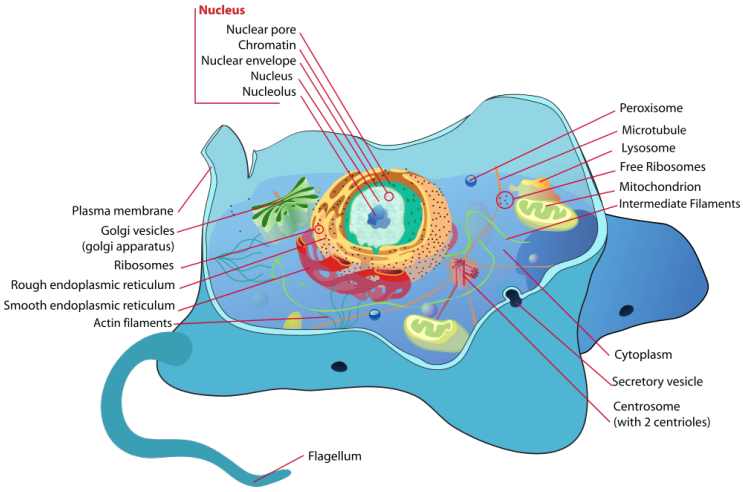


Figure 2.2: Animal eukaryotic cell diagram. Image source [Rui].

2.1.1 Transcription process

Transcription is the process in which the instructions encoded in a gene are copied from the DNA inside the nucleus, to produce a RNA transcript called messenger RNA. The Transcription process has two broad 2 main steps; 1) The process in which the gene is copied from the DNA into a pre-processed version of the messenger RNA (mRNA) called pre-messenger RNA (pre-mRNA); 2) The RNA Splicing, which is the process where the pre-mRNA is transformed into a mature mRNA strand. These two processes happen inside the cell nucleus.

Step 1, Pre-messenger RNA synthesis

The pre-mRNA creation has 3 main processes[JD+13] and is illustrated in figure 2.3:

1. **Initiation**. This step initiate the transcription process. It happen when an *enzyme*[3] of RNA polymerase binds to a region of the target gene called *the promoter*. This indicate the DNA to unwind, so the RNA polymerase can read the DNA bases in one of its strands and create a molecule of pre-mRNA.

2. **Elongation**. Elongation is the process in which *nucleotides*[4] are added to the pre-mRNA strand. The RNA polymerase enzyme read the unwound DNA strand

[3] An enzyme is a proteins that act as biological catalysts to accelerate chemical reactions.

[4] Nucleotides are the building block of nucleic acids. A nucleotide consists of a sugar molecule bound to a phosphate group and a nitrogen-containing base. RNA and DNA are polymers made of long chains of nucleotides.

and synthesize the pre-mRNA molecule.

3. **Termination**. Termination ends the pre-mRNA synthesis. It happens when the RNA polymerase enzyme identifies a termination sequence in the gene, and detaches from the unwound DNA.

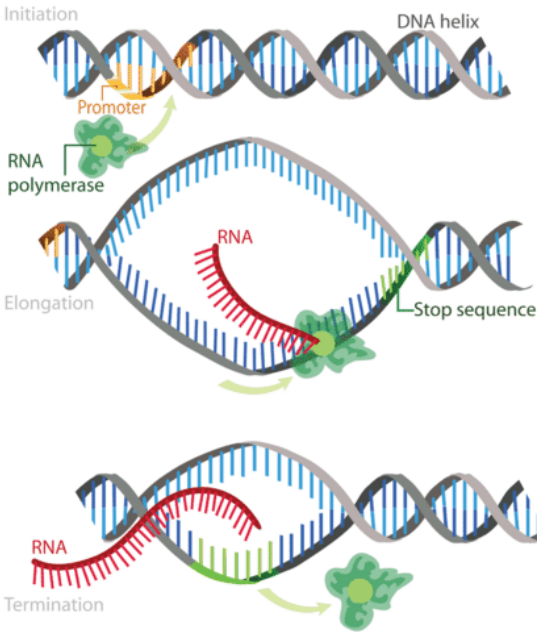


Figure 2.3: The three main steps of the pre-mRNA synthesis: initiation, elongation, and termination. Image source [Vil].

Step 2, Pre-messenger RNA splicing

In this process the pre-mRNA is transformed into a mature mRNA strand by removing non-relevant (non-coding) sections of it. During Splicing, *introns*[5] are removed and *exons*[6] are joined together [Ber+15]. Besides this, a cap and a tail are added to the spliced pre-mRNA strand to turn it into a mature mRNA. The splicing process takes place within subnuclear structures called Nuclear Speckles (NS) (also known as *Splicing Speckles*) [SL11]. Figure 2.4 illustrates the pre-mRNA splicing process.

[5]Introns are nucleotide sequences within a gene that are non-coding regions of an RNA transcript, and that are removed by the splicing process before translation.

[6]Exons are coding sections of an RNA transcript that can be translated into proteins.

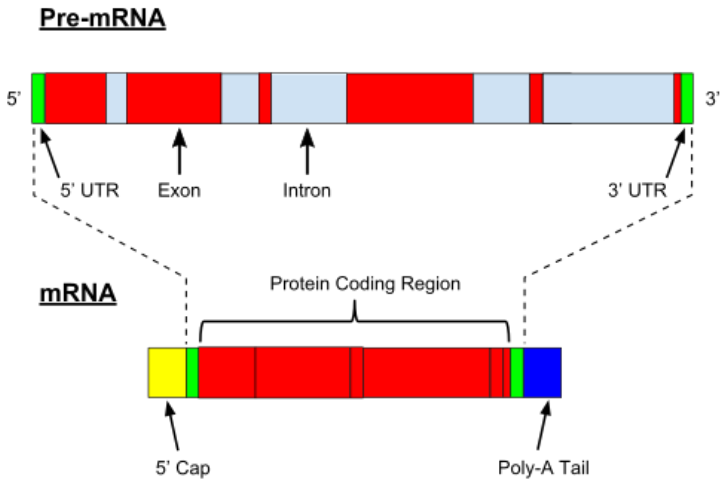


Figure 2.4: Pre-messenger RNA splicing process. A pre-mRNA strand (top) is turned into a mature mRNA strand (bottom). Image source [Wik21].

Transcription Rate

Transcription Rate (TR) contemplates the two steps we already explained; 1) the nascent transcription rate, which measures the in situ mRNA produced by the RNA polymerase enzyme (pre-mRNA), and 2) the rate of synthesis of mature mRNA (pre-mRNA splicing), which measures the contribution of transcription to the mRNA concentration [PO+13].

Accordingly to Pérez-Ortín et al. [PO+13], we can define the change in the mature mRNA concentration ([mRNA]), as the number of mature mRNA molecules being synthesized per unit of time minus a degradation factor

$$\frac{d[mRNA]}{dt} = SR - k_d[mRNA] \tag{2.1}$$

where $[mRNA]$ is the mature mRNA concentration in the cell nucleus, SR the mature mRNA synthesis rate and k_d the degradation rate.

However, the objective of this work is not to model the change in TR across time. Instead, we are only interested in estimating the TR of a cell, given a snapshot of it at a specific moment of time. For this reason, in this work we will understand TR as the amount of nascent mRNA in a given period of time [Wan+93].

2.2 Machine learning

Artificial Neural Networks (ANNs) are universal approximators widely used in the field of Machine Learning (ML) and an important part of this work. This is a very broad subject and there are entire books that cover this in detail, like [GBC16] or [Bis06]. However, in this section we will give a small introduction to ANNs and Convolutional Neural Networks (CNNs), which are a type of ANN that were specifically designed to deal with data in the form of images.

Before defining what exactly is a ANN, lest first recall the definition of machine learning. We refer as ML to the group of algorithms that automatically improve (learn) through experience. Among this algorithms, we could say that there are three main classes (which depend on the kind of experience we provide):

- **Supervised Learning**: The experience is given in the form of input and output examples, and the goal is to learn a general rule that maps inputs to outputs.
- **Unsupervised Learning**: The experience is given in the form of data (no outputs provided) and the goal is to discover hidden patterns in data.
- **Reinforcement Learning**: No experience (data) is given, instead a dynamic *environment* is provided and an *agent* must learn how to interact with it in order to achieve a goal.

An ANN can be used in any of the 3 kinds of learning algorithms listed above.

However, recall that the objective of this work is to approximate a function (in this case a CNN), such that when it is fed with images of a cell nucleus (input data), it predicts the corresponding Transcription Rate (TR) (output data) Therefore, we are dealing with a *supervised learning* task.

Before explaining what a CNN is, let us first introduce and explain ANN in general.

2.2.1 Artificial neural networks

Roughly speaking, an Artificial Neural Network (ANN) is a non-linear function $f : \mathbb{R}^D \to \mathbb{R}^L$, that maps an input $\boldsymbol{x} \in \mathbb{R}^D$ with an output $\boldsymbol{y} \in \mathbb{R}^L$. Of course, to consider f as a ANN, f must have a specific form that will be address later. However, for the sake of this explanation, let us start by defining a simple function as follow

$$\begin{aligned} f(\boldsymbol{x}, \boldsymbol{w}) &:= h\left(w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\boldsymbol{x})\right) \\ &= h(\boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})) \\ &:= h(z) \end{aligned} \tag{2.2}$$

where $\boldsymbol{\phi} : \mathbb{R}^{D+1} \to \mathbb{R}^M$ is an element-wise function, with $\phi_0 := 1$, know as *basis function*, $h : \mathbb{R} \to \mathbb{R}$ is a function know as *activation function* and $\boldsymbol{w} \in \mathbb{R}^M$ is the parameter vector. The parameters w_j, with $j \in \{1, \dots M-1\}$ are known as *weights*, while the parameter w_0 is know as *bias*.

Then, an ANN is composition of functions of the same form as 2.2, with non-linear *activation functions*, and where the basis functions are also of the same form as 2.2 [Bis06]

$$F(\boldsymbol{x}, \boldsymbol{W}) := h_K(\boldsymbol{w}_K^T h_{K-1}(\boldsymbol{w}_{K-1}^T \dots h_0(\boldsymbol{w}_0^T \boldsymbol{x}) \dots)) \tag{2.3}$$

The subscript in the parameter vectors $\boldsymbol{w}_k$ and the activation functions h_k, with $k \in \{0, \dots, K\}$, of 2.3 represents the depth of the layers. Note that unlike the other layers, the base function of the *input layer* (k = 0) is the identity function. Furthermore, the activation function of the *output layer* h_K does not necessarily have to be non-linear. Instead, it is chosen based on the type of function we want to approximate. In our case, since we have a regression problem (predicting Transcription Rate (TR)), h_K is chosen as the identity function.

There are different non-linear activation functions that can be chosen for the hidden units. However, all the models showed in this work use the Rectified Linear Unit (ReLU)

$$ReLU := max\{0, x\} \tag{2.4}$$

Figure 2.5 shows the ReLU activation function.

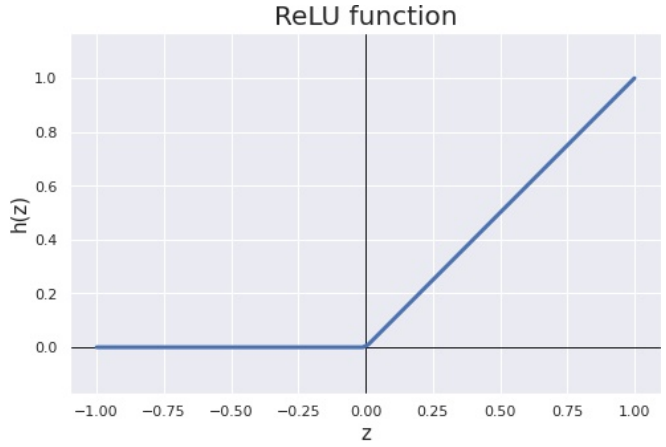


Figure 2.5: ReLU activation function.

Figure 2.6 shows a graphical representation of a ANN. The circles represent the activation function applied to what is inside it. Black colored circles represent the

identity function, red colored circles the non-linear activation function for the hidden layers, while green any function for the output layer that suits the problem we want to solve. Note that values inside the circles of the hidden and output layers z_i^k, for $k \in \{0, \dots, K\}$ and i representing one of the units of the k layer, are the output of a function of the same form as 2.2. The lines connecting the circles represent the weights and biases corresponding to each layer $\boldsymbol{W}_k$, for $k \in \{0, \dots, K\}$. The circles in the *hidden layers* are known as *hidden units*.

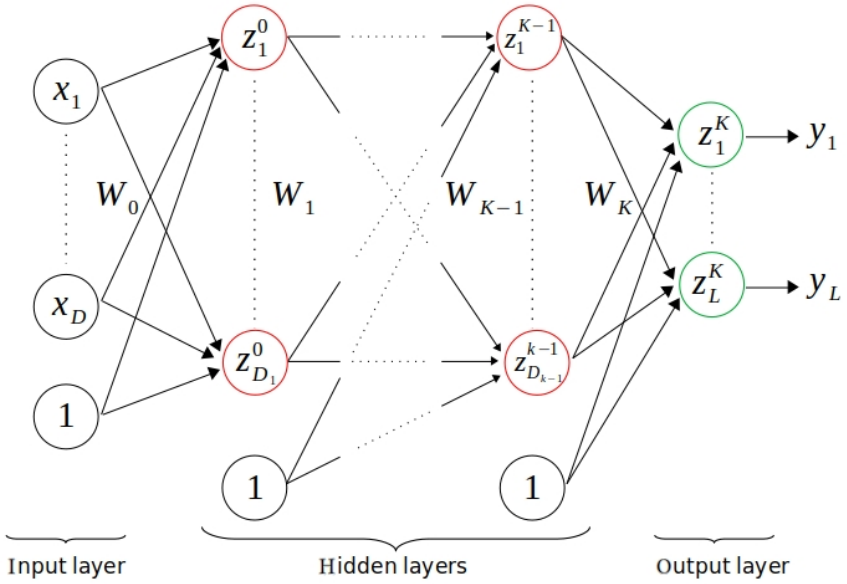


Figure 2.6: Graphical representation of an ANN. The color of the circles represents the type of activation function. Black means the identity, red a non-linear function for the hidden layers and green any function for the output layer.

Strictly speaking, equation 2.3 and figure 2.6 represent a *fully connected feedforward neural network*. However, in this work we will refer to it just as ANN, which in some literature is also known as Multilayer Perceptron (MLP). Also, hidden layers are also known as *Dense layers*.

Update rule

Sow far we have introduced the general form an ANN must have. Moreover, equation 2.3 shows that an ANN is simply a non-linear function controlled by a set of adjustable parameters $\boldsymbol{W}$. Therefor the question is, how can we approximate this parameters?

Recall that we are dealing with a supervised learning problem, which means that we can use both the input data (images of cell nucleus, $\boldsymbol{X}$) and the output data (the TRs,

$\boldsymbol{Y}$) to approximate $\boldsymbol{W}$. Therefore, we can fed the ANN with $\boldsymbol{X}$, and then measure its performance by comparing its output $\hat{\boldsymbol{Y}}$ against the true values $\boldsymbol{Y}$.

This comparison is made by means of a *loss function* $\mathcal{L}$ that must be chosen beforehand. The choice of $\mathcal{L}$ depends mainly on the type of problem you are solving (regression, classification, etc.). However, even for each type, there are many different options. For now, let us just say that $\mathcal{L}$ should return high values when $\hat{\boldsymbol{Y}}$ is far from the true values $\boldsymbol{Y}$, and low when they are close.

Then, we can fit the values of $\boldsymbol{W}$, by minimizing the loss function $\mathcal{L}$ each time the model is fed with an input value $\boldsymbol{x}$. Since the gradient of $\mathcal{L}$ with respect to $\boldsymbol{W}$ (i.e., $\nabla_{\boldsymbol{W}}\mathcal{L}$) returns the direction in which the loss function grows the fastest, then we choose $-\nabla_{\boldsymbol{W}}\mathcal{L}$ as the direction of our update rule

$$\boldsymbol{W}_{new} = \boldsymbol{W}_{old} - \alpha\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_{old}) \tag{2.5}$$

where $\alpha \in \mathbb{R}^+$ (known as *learning rate*) controls how much we move in the direction of $-\nabla_{\boldsymbol{W}}\mathcal{L}(\boldsymbol{W}_{old})$ on every step.

The iterative method in which 2.5 is applied over elements of $\boldsymbol{X}$ to optimize $\boldsymbol{W}$ is known as Gradient Descent (GD) [Bis06]. However, in practice 2.5 is not applied for a single element of $\boldsymbol{X}$ every time, but to a random subset of $\boldsymbol{X}$ (known as a *Batch*) instead. The number of elements in batch is fixed over all the iteration (training), and is an hyperparameter known as *Batch Size* bs[7]. As a rule of thumb, bs should be no less than 30 (for the selected sample to be representative of $\boldsymbol{X}$). In practice bs is usually chosen as a power of 2. This process is known as Stochastic Gradient Descent (SGD) and computationally is less expensive than GD.

However, GD (SGD) has a downside, the choice of its hyperparameter α (learning rate). In practice, it has been shown that the correct choice of α is essential to train an ANN successfully. Therefore, other algorithms (*optimizers*) have been proposed to mitigate this problem. The revision of these optimizers is out of the scope to this work. However, all of them follow the same idea proposed by GD. For example, instead of having a fixed learning rate α as in GD, the Adaptive Moment Estimation (Adam) optimizer adapts its learning rate dynamically during training depending on the mean and variance of the loss function [KB14].

Back propagation

Nevertheless, there is still one question that needs to be answered, which is how to efficiently calculate the derivative of the loss function ($\nabla_{\boldsymbol{W}}\mathcal{L}$) with respect to all the parameters of the ANN. The answer to this is through an algorithm called *backpropagation*, which is performed during the *training process*. Again, there is a lot

[7] Normally the training data is separated in disjoint batches, which means that it could happen that last batch to be smaller than the selected bs.

of literature that explains this in depth(for instance [GBC16] or [Bis06]). Therefor, here we will just provide the intuition behind it.

Recall that $\mathcal{L}$ is a function of the true values y and $\hat{y}$ i.e., $\mathcal{L}(y, \hat{y})$. Also from equation 2.3 and figure 2.6 note that

$$\begin{aligned} y &:= F(\boldsymbol{x}, \boldsymbol{W}) \\ &= h_K(\boldsymbol{z}^K) \\ &= h_K(\boldsymbol{W}_K^T h_{K-1}(\boldsymbol{z}^{K-1})) \end{aligned} \tag{2.6}$$

and therefore

$$\begin{aligned} \nabla_{\boldsymbol{W}_K} \mathcal{L} &= \frac{\partial \mathcal{L}}{\partial \boldsymbol{W}_K} \\ &= \frac{\partial \mathcal{L}}{\partial \hat{y}} \frac{\partial \hat{y}}{\partial \boldsymbol{z}^{\boldsymbol{K}}} \frac{\partial \boldsymbol{z}^{\boldsymbol{K}}}{\partial \boldsymbol{W}_K} \end{aligned} \tag{2.7}$$

which is just the product of the derivative of the loss function w.r.t. $\hat{y}$ (i.e., $\frac{\partial \mathcal{L}}{\partial \hat{y}}$), the derivative of the activation function of the output layer w.r.t the argument of the last layer (i.e., $\frac{\partial \hat{y}}{\partial \boldsymbol{z}^{\boldsymbol{K}}}$) and the output of the layer $K-1$ (i.e., $\frac{\partial \boldsymbol{z}^{\boldsymbol{K}}}{\partial \boldsymbol{W}_K} = h_{K-1}(\boldsymbol{z}^{\boldsymbol{K-1}})$).

Note that we can easily compute the gradient of $\mathcal{L}$ w.r.t deeper parameters $\boldsymbol{W}_k$ (for $k \in \{0, \ldots, K-1\}$), just by extending 2.6 and 2.7.

This shows how by means of the *chain rule*[8], the backpropagation algorithm can compute the gradient of the loss function w.r.t. a specific parameter, just by multiplying the derivative of the loss function, the derivative of the activation functions and some values computed during the evaluation of the ANN.

Model development

The properties of ANNs have been studied extensively before ([Cyb89], [HSW89], [Fun89]) and established in the *Universal approximation theorem*

Theorem 2.1 (Universal approximation theorem)
An MLP with a linear output layer and one hidden layer can approximate any continuous function defined over a closed and bounded subset of $\mathbb{R}^D$, under mild assumptions on the activation function (squashing activation function) and given the number of hidden units is large enough.

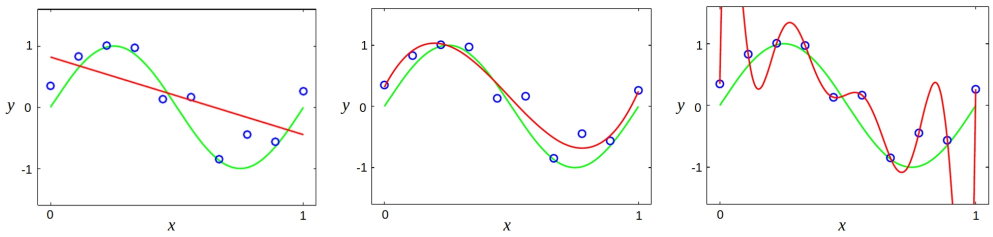
[8] $(f \circ g)' = (f \circ g) \cdot g'$, or equivalently $h'(x) = f'(g(x))g'(x)$, for $h(x) := f(g(x))$.

For this reason ANN are known as *universal approximators*, since they are able to approximate any continuous function on a compact[9] input domain with an arbitrary accuracy [Bis06].

These means that, as long as a ANN has a sufficiently large number of hidden units, the loss function can be reduced as much as desired. However, this nice property can also lead to an unwanted one known as *overfitting*. Intuitively this means that the ANN *memorize* the data used to train it (low error/bias), and therefore it is not able to perform (or *generalize*) well when it is fed with new data (high error/bias and variance). This happens mainly when the ANN is optimized/fed too many times with the same data.

On the other hand, *underfitting* means that the ANN performs poorly on both new data and data used to train the network (high bias and low variance). This usually happens when the training time is insufficient or the ANN is not complex enough (too few hidden units and/or layers).

Figure 2.7 shows synthetic data (blue circles), generated from a sine function (green line) and random noise sampled from a normal distribution. The red line in figure 2.7a represents a fitted model with high bias and low variance (underfitting), while in figure 2.7c a model with low bias and high variance (overfitting). The red line in figure 2.7b, represents a model with low bias and variance (good fit and good generalization).



(a) Underfitted model. **(b)** Model with good fit and generalization. **(c)** Overfitted model.

Figure 2.7: Representation of a model (red line) with underfitting a), good fit b) and overffiting c), trained over synthetic data (blue small circles). The synthetic data was generating by adding random noise to a sine function (green line) on the interval $[0, 1]$. Image source [Bis06].

In practice, we seek to fit models that has low bias and low variance (i.e., good accuracy and good generalization). Therefore, to prevent overfitting we split the data into 3 different sets; *training*, *validation* and *test*, and train the model using only the first set. Then, during model training, we measure how well the model is generalizing

[9] A set A in a metric space is said to be *compact* if it is close (i.e., it contain all its limit points) and bounded (i.e., all its points lie within some fixed distance of each other) [BS00].

by comparing the value of the loss function when it is evaluated in the training and validation set [10]. During the model development, the *test* set is never evaluated and is only used at the end, to report the model performance. This methodology is shown in figure 2.9

Figure 2.8 shows this *bias–variance tradeoff* between training and validation set. In practice, multiple versions of the model are saved during training and then the one with the lowest validation error is chosen (red dot on figure 2.8).

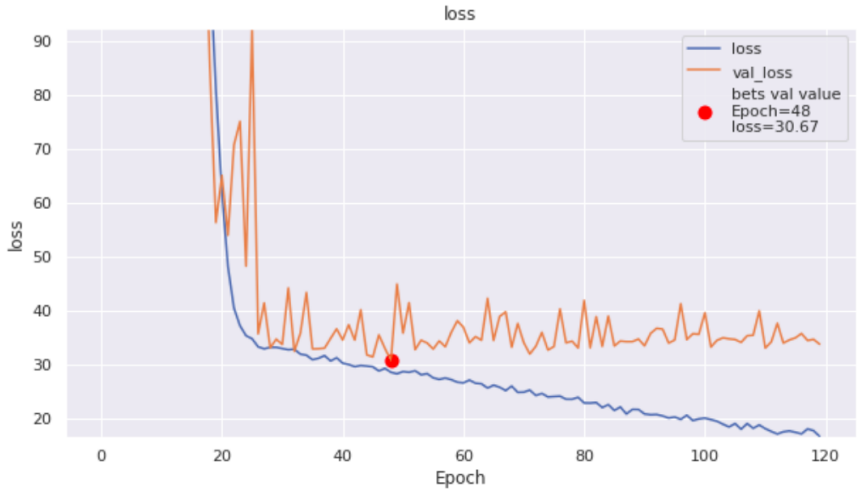


Figure 2.8: Bias–variance tradeoff. In orange (respectively blue) the loss function curve when it is evaluated in the validation (respectively training) set. The red dot shows the lowest loss for the validation set.

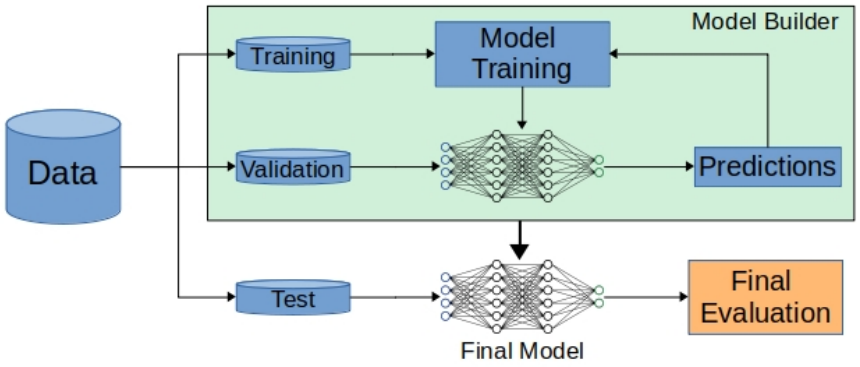


Figure 2.9: Model development methodology.

[10] This is usually known as the *bias–variance tradeoff*.

The methodology shown in figure 2.9 is also used to optimize the hyperparameters of the model, like the number hidden units/layers or the activation function of the hidden layers.

Batch Normalization

However, overfitting is not the only problem we may encounter when training an ANN. Training ANN with several layers can be complicates, since the distribution of the data can change from layer to layer. This means that the input and output distribution of a layer will not necessarily be the same. It has been empirically proven that this can affect the training performance, since it require the use of lower learning rates [IS15]. This can also lead to *saturation*[11] of the activation functions, so a more careful initialization of the ANN parameters is required. To address this problem Ioffe et al. [IS15] proposed to normalize the layer inputs.

Roughly speaking, batch normalization consist of two main steps; 1) the standardization of the layer input and 2) the normalization of the standardized data. For the first step the layer input is standardized using parameters extracted from the *batch*

$$\boldsymbol{z}'_k := \frac{\boldsymbol{z}_k - \boldsymbol{\mu}_k}{\sqrt{\boldsymbol{\sigma}_k^2 - \epsilon}} \tag{2.8}$$

with

$$\begin{aligned} \boldsymbol{\mu}_k &= \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{z}_k \\ \boldsymbol{\sigma}_k^2 &= \frac{1}{M}\sum_{m=1}^{M} (\boldsymbol{z}_k - \boldsymbol{\mu}_k)^2 \end{aligned} \tag{2.9}$$

where M is the *Batch* size and k, with $k \in \{0 \dots K\}$, denotes the layer.

Note that for each layer k we have different normalization parameters $\boldsymbol{\mu}_k$ and $\boldsymbol{\sigma}_k$. Moreover, this normalization parameters are vectors of the same shape as the layer size (i.e., one pair of normalization parameters per unit/neuron).

The second step in batch normalization consist on normalizing the standardized data $\boldsymbol{z}'_k$ using parameters $\boldsymbol{\gamma}_k$ and $\boldsymbol{\beta}_k$ learned during training

$$\widetilde{\boldsymbol{z}}_k := \boldsymbol{\gamma}_k \odot \boldsymbol{z}'_k + \boldsymbol{\beta}_k \tag{2.10}$$

where $\odot$ denotes *element-wise* multiplication. At the beginning of the training $\boldsymbol{\gamma}_k = 1$ and $\boldsymbol{\beta}_k = 0$ are used for all the layers and units.

[11] *Saturation* is a commonly used term to refer to the situation when the evaluation of a "squashing" function returns values close to some of its horizontal asymptotes most of the time. Remember that these "squash" functions (like *Sigmoid* or *tanh*) compress the real line $(-\inf, \inf)$ into an interval of finite length (a, b).

During training, the normalization parameters of each epoch are stored, so the average ($\bar{\gamma}_k$ and $\bar{\beta}_k$) can be used during evaluation (when the model is not training).

Residual Block V2

As already mentioned, the *Universal approximation theorem* guarantees that the training error can be reduced by adding more layer to an ANN. However, in practice it is not that simple. As we add layers to an ANN, the training becomes more unstable and difficult as we can face vanishing or exploding gradients (when the value of the gradients become very close to 0 or inf respectively during back propagation). To overcome this problem, He et al. ([He+15] and [He+16]) proposed the *residual blocks*, which have been empirically shown to make deep ANN training more stable. The core idea of residual blocks is to reformulate the layers as *learning residual functions* with reference to the layer inputs, by adding an *identity connection*. Then, if a layer is not longer beneficial to the ANN (e.g. in case of gradient vanishing), the ANN can just "skip" it. Figure 2.10 shows a diagram of the second version of a residual block [He+16].

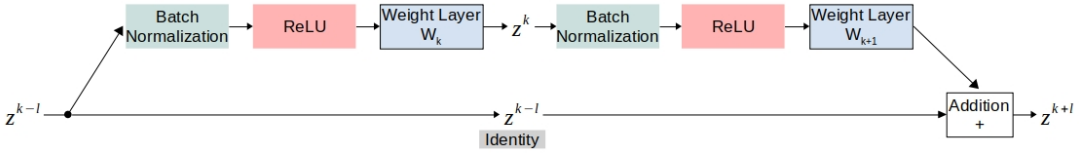


Figure 2.10: Residual block V2.

2.2.2 Convolutional neural networks

So far we have explained how Artificial Neural Networks (ANNs) works assuming that we feed them with vectors of fixed length. Even though we could take a multichannel image and transform it into a vector, in practice this would be computationally very expensive. For instance, assuming that we have a 3 channel image of size 224 by 224, this would result into an input vector of length $3 \cdot 224 \cdot 224 = 150'528$. Then, if the first layer of our network has 100 units, this would mean more than 15 millions of parameters only for the first layer. Furthermore, the transformation of our image into a vector would mean a loss of spatial information. This means that the ANN would not be able to capture or use the spatial relationship between pixels and shapes within the image.

A Convolutional Neural Network (CNN) is a type of ANN widely used to analyze data in the form of images. The intuition behind a CNN is that instead of just looking at an image and trying to predict the target value directly, first learn some *features* within the image, and then make the predict base on this features. To achieve this, CNNs mainly use *convolution* and *pooling* layers.

Convolution layer

The only difference a

A convolution layer is very similar to a regular layer described in section 2.2.1. Basically, they only differ in the way the layer input is multiplied by the the layer weights. Recall that in a regular layer, the input of a unit is the dot product between the layer input and its corresponding weight vector (i.e., $z = \boldsymbol{w}^T \boldsymbol{x}$). This means that for each element in the input vector $\boldsymbol{x}$, there is a corresponding element in the weight vector $\boldsymbol{w}$. However, for a convolution layer this is not the case. Convolution layers are based on the shared-weight architecture of the convolution *kernels* or *filters* that slide along the input and returns a translation known as *feature maps* [Zha+88]. This means that the *kernels* weights will be used for multiple elements of the layer input. Figure 2.11 shows the convolution process with a 2 by 2 kernel over a RGB image (3 channels) of size 4 by 4. Each entrance of the returned feature map z_i is the dot product between the kernel weights $\boldsymbol{w}$ and the $\boldsymbol{x}_i - th$ chunk of the image.

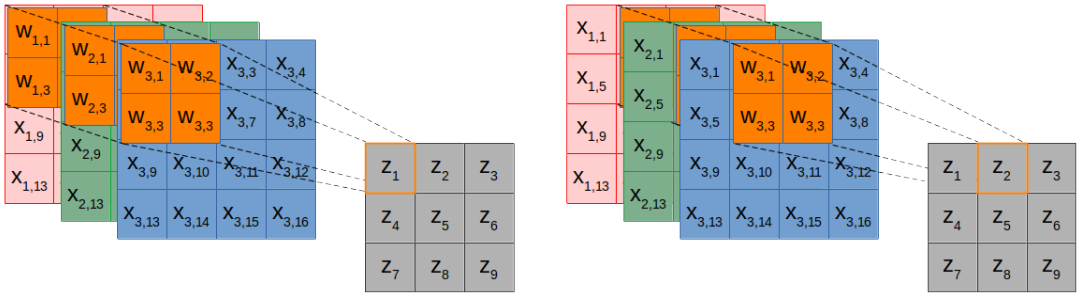


Figure 2.11: Convolution process steps. In red, green and blue the input image, in orange the convolution kernel (size 2 by 2 and stride of 1) and in gray the convolution output (feature map).

Mathematically this looks as follow

$$z_i = \boldsymbol{w}^T \boldsymbol{x}_i + b \tag{2.11}$$

where $\boldsymbol{w} \in \mathbb{R}^{2\times2\times3}$, $\boldsymbol{x_i} \in \mathbb{R}^{2\times2\times3}$ and $b \in \mathbb{R}$ is the bias (not shown in the images).

Like the kernel size, the number of pixels we shift the kernel each time along side the input (*Stride*) is also a hyperparameter of convolution layers. IN figure 2.11, the stride size is 1.

Figure 2.11 also shows that size (width and height) of the returned feature map is smaller than the input image. If we want to keep the input and output size the same (*Same convolution*), then we must add zeros at the edges of the input features (zero-padding). This is shown on figure 2.12.

So far we have seen that a convolution projects a multi-channel input feature (image)

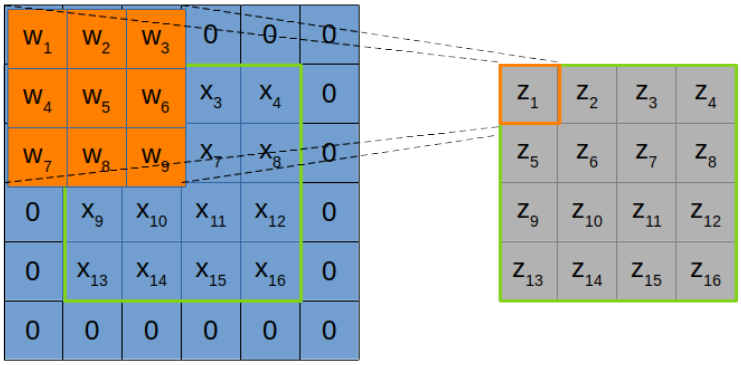


Figure 2.12: Convolution with padding. In blue a single-channel input features, in orange the convolution kernel (size 3 by 3 and stride of 1) and in gray the convolution output (feature map).

into a single-channel feature map. Therefore, if we want our output feature map to have n channels, then our convolution must have n different kernels.

Normally, a non-linear activation function is applied to the output of convolution layers (and normally also after batch normalization) to enable the CNN to learn non-linear relations.

Pooling layer

Unlike convolution layers, the goal of Pooling layers is to reduce the feature image (height and width, but not depth) rather than learn features. However, Pooling layers work in a similar way to convolution layers in the way that they also slide a kernel along the input. However, in this case the kernel works independently on each feature map (that is, each channel) and has no weights to learn. This means that the pooling layers maintain the same number of input and output channels. There are several ways to do this downsampling, but the most common are Max Polling and Average Polling. As the name suggests, Average pooling shrinks the feature image by averaging sections of it, while Max pooling takes the maximum value. Figure 2.13 shows an example of a max and average pooling layer on a single-channel feature image using a 2 by 2 kernel and a stride of 2.

Normally, Pooling layers are applied over the output of the activation functions.

Global Average Pooling layer

As we mentioned at the beginning of this section, the idea of a CNN is to first learn the features within the input images and then make a prediction based on these features. To do this, the *Global Average Pooling layer* transforms the channels of the last feature

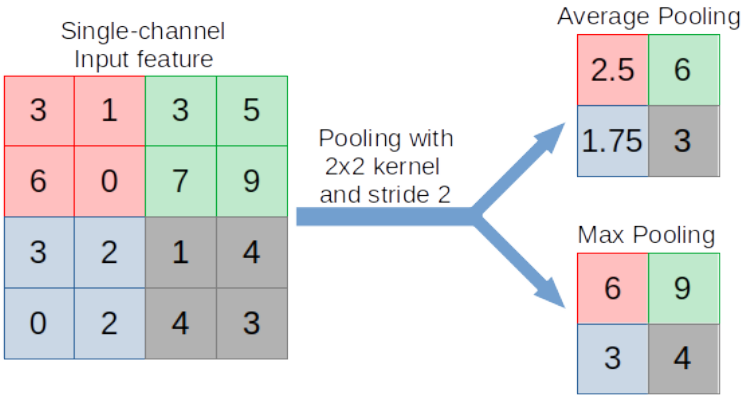


Figure 2.13: Max and average pooling with a 2 by 2 kernel and stride 2. The color denotes the kernel position.

map into a vector (by averaging each of its channels), so that this can be used as input in a regular ANN to make the final prediction. Figure 2.14 shows an example of this, when it is applied into a feature map with 7 channels.

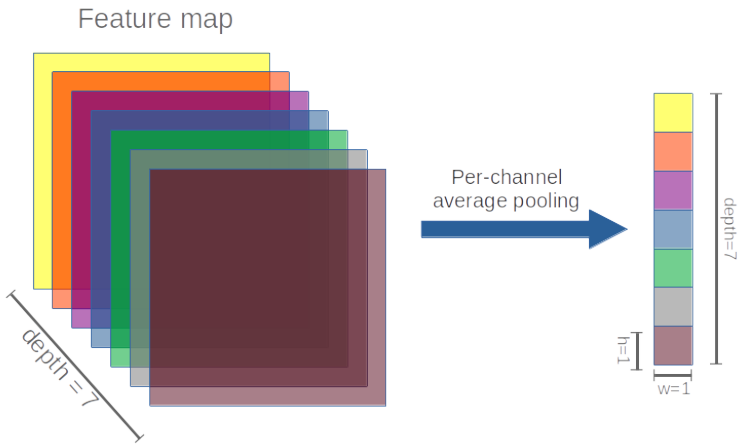


Figure 2.14: Global Average Pooling layer.

Inception module

Recall that a convolution layer is meant to learn features from a 3D object with 2 spatial dimensions (width and height) and a channel dimension. This means that each kernel in the convolution needs to learn simultaneously cross-channel and spatial correlations. The intuition behind the *Inception module* is to improve this process

by separating this two tasks, so that the cross-channel correlations and the spatial correlations can be learned separately and independently [Cho17].

A normal inception model looks at the cross-channel correlations first through a set of 3 or 4 *pointwise convolutions*[12], and then learns the spacial information in the downsampled feature image (in depth, not height and width), by means of regular convolution (usually with 3 by 3 or 5 by 5 kernels). Figure 2.15 shows a diagram of an Inveption V3 module.

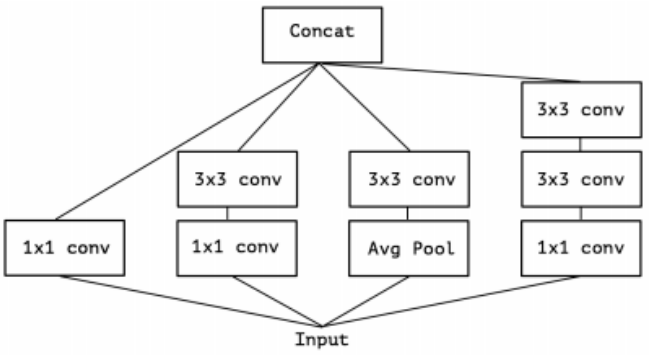


Figure 2.15: A regular Inception module (Inveption V3). Image source [Cho17].

François Chollet [Cho17], used the inception module as reference to propose the *depthwise separable convolution*, which is something between a normal convolution and a normal convolution combined/followed by a pointwise convolution. Figure 2.16 shows an *extreme* version of the inception module shown in figure 2.15. The *depthwise separable convolution* is very similar to the one shown in figure 2.16, the only difference is that the pointwise convolution is applied before the 3 by 3 convolutions instead of after.

Even though the *depthwise separable convolution* is a simplified version of the inception module, the idea and motivation behind it is the same. The *depthwise separable convolution*, and the residual block, are the main components of the *Xception* architecture [Cho17].

2.3 Interpretability methods

In recent years, Deep Neural Networks (DNNs) have been used to solve a wide variety of problems and gained popularity. Amazing results such as those achieved by Deep Mind's Alpha Fold team, have shown the great potential DNN has to solve complex problems [Sen+20]. However, the difficulty to interpret DNNs has become one of the

[12] A *pointwise convolution* is a convolution with 1 by 1 kernels and stride 1.

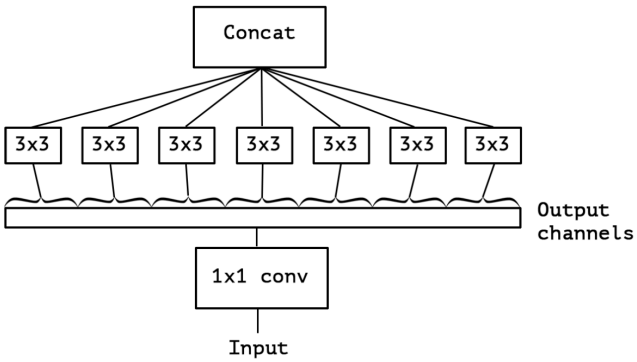


Figure 2.16: An extreme version of our Inception module. Image source [Cho17].

main obstacles to their acceptance in applications where the interpretability of the model is necessary.

To understand how the DNNs predict the Transcription Rate (TR) of a cell, we use *Attribution Methods*. This methods are meant to measure how much each component of the input image contributes to the model's prediction by creating a *Score Map* (also known as *Importance Map, Sensitivity Map* or *Saliency Map*) of the same shape as the model's input. In particular, in this work we use a combination between Integrated Gradient (IG) [STY17] and VarGrad (VG) [Ade+20] as attribution method. In general we will denote attribution method as ϕ.

Attribution methods are not only used to interpret black-box models like DNN, the can also be used to debug models or as a sanity check to validate that the model base its prediction on the relevant features of the input.

In our case, this interpretability techniques will show us which parts of the cell image are relevant for the prediction of the TR. However, this will not just help us to interpret the results of the model, this also have the potential to help us understand unknown cellular processes.

2.3.1 Integrated gradients

Integrated Gradient (IG) is an interpretability technique (attribution method) proposed by Sundararajan et al. [STY17], aimed to assign an importance to the input features (in our case pixels from a cell image) with respect to the model prediction. The attribution problem have been studied before in other papers [Bae+10], [SVZ13], [Shr+16], [Bin+16] and [Spr+14].

In our case, we seek to predict Transcription Rate (TR) given a cell image $x \in \mathbb{R}^{d \times d \times c}$, where d is the height and width of the image and c is the number of channels. Therefore, our Deep Neural Network (DNN) would be a function $f : \mathbb{R}^{d \times d \times c} \to \mathbb{R}$ and an attribution method should be a function $\phi : \mathbb{R}^{d \times d \times c} \to \mathbb{R}^{d \times d \times c}$ having an input and

output of the same shape as the model's input image.

Early interpretability methods only use gradients to assign importance to each input feature

$$\begin{aligned}\phi(f,x) &:= \nabla f(x) \\ &= \frac{\partial f}{\partial x}\end{aligned} \tag{2.12}$$

Mathematically speaking, $\phi_i(f,x)$ assign an importance score to the pixel i (out of the $d \times d \times c$ there are), representing how much it adds or subtract from the model output. However, this score maps have some drawback when they are used to interpret deep neural networks [SLL20]. Recall that the gradient with respect to the input indicate us the pixels that have the steepest local slope with respect to the model's output. This means that it only describes local changes in the input, and not the whole prediction model. Another mayor problem is saturation[13]. As the model learns the relationship between an input image and its TR, the gradient of the most important pixels will approximate to 0, i.e. the pixel's gradient saturates.

To overcome this problems, Sundararajan et al. proposed Integrated Gradient (IG) as an attribution method, where the importance of the input feature i is defined as follow

$$\phi_i^{IG}(f,x,x') := (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha \tag{2.13}$$

Intuitively speaking, IG accumulates the input gradient when it goes from a baseline x', which should represents *absence* of information, to the actual input image x. With this, we avoid losing information about relevant pixels for the model's prediction in the importance map, even if they saturate eventually. Figure 2.17 shows an example of the image progression fed into IG. Note that the amount of information in the images is parameterized by $\alpha \in [0,1]$, and that the *absence* of information is interpreted as a black image.

For a better understanding, we can divide the IG definition as follow

$$\phi_i^{IG}(f,x,x') := \overbrace{(x_i - x_i')}^{\text{Difference from baseline}} \underbrace{\int_{\alpha=0}^{1}}_{\text{From baseline to input...}} \overbrace{\frac{\partial f(x' + \alpha(x - x'))}{\partial x_i} d\alpha}^{\text{...accumulate local gradients}} \tag{2.14}$$

The integral in equation 2.14 accumulate the gradients for the interpolated images $x' + \alpha(x - x'))$ between the baseline x' and the image x. On the other hand, the

[13] In the context of artificial neural networks, a neuron is said to be saturated when the predominant output value of a neuron is close to the asymptotic ends of the bounded activation function. This behavior can potentially damage the learning capacity of a neural network.

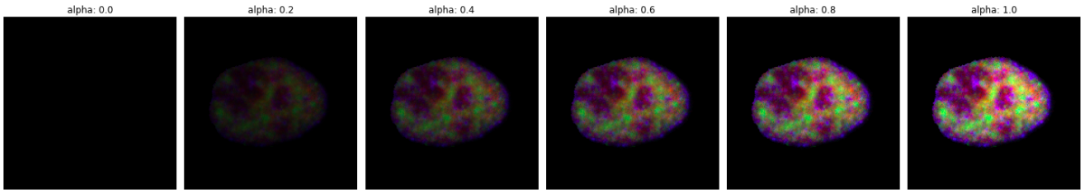


Figure 2.17: Progression from an image with no information (back image) to a normal one parameterized by α.

difference $(x_i - x_i')$ outside the integral comes from the chain rule and the fact that we are interested in integrating over the path between the baseline and the image.

IG is very simple and easy to implement, since it does not require any modification to the model and it only require some calls to the gradient operator.

The IG satisfy several properties and axioms that are addressed in detail in the paper. However, there is one axiom satisfied by IG that is of special importance for us, *completeness*. Completeness means that the value of the summed attributes will be equal to difference between the model's output when it is evaluated at the image and the model's output when it is evaluated at the baseline

$$\sum_i \phi(f, x, x')^{IG} = f(x) - f(x') \tag{2.15}$$

In practice, computing the analytic expression for the integral in equation 2.13 would be complicated, and in some cases unfeasible. However, luckily we can numerically approximate $\phi(f, x, x')^{IG}$ using a Riemann sum

$$\phi_i^{Approx\ IG}(f, x, x', m) := (x_i - x_i') \sum_{k=1}^{m} \frac{\partial f(x' + \frac{k}{m}(x - x'))}{\partial x_i} \frac{1}{m} \tag{2.16}$$

where m is number of steps for the Riemann sum approximation.

This is when the completeness axiom comes into scene, which is a good value for the parameter m? 10, 100, 500? To answer this question, we can simply apply the completeness axiom as a sanity check for the election of m. If m is good enough, then the value of $\sum_i \phi_i^{Approx\ IG}(f, x, x', m)$ should be close to $f(x) - f(x')$, or equivalently, the value of $|\sum_i \phi_i^{Approx\ IG}(f, x, x', m) - (f(x) - f(x'))|$ should be close to 0.

Figures 2.18b and 2.18c show a comparison between the gradient of a model output with respect to a cell image, and the IG. One can see that either for score maps computed using IG or vanilla gradients, the output is noisy and diffuse.

2.3.2 VarGrad

As we can see in figure 2.18c, Integrated Gradient (IG) attribution maps can be noisy and diffuse. To improve their empirical quality, Smilkov et al. [Smi+17] proposed

SmoothGrad (SG), which tends to reduce noise in practice and can be combined with other attribution map algorithms (like IG). The idea behind SG is pretty simple, given an input image x, you create a sample of similar images by adding noise, then compute the attribution map for each one of them using the algorithm you prefer (in our case IG), and take the average of the attribution maps. Although Smilkov et al. do not provide a mathematical proof of why SG reduce noise in score maps, they provide a conjecture and empirical evidence. For this work we use a slightly different version called VarGrad (VG), proposed by Adebayo et al. [Ade+18] but inspired by SG, which takes the variance of the attribution maps instead of the mean. The reason for this choice is that Seo et al. [Seo+18] analyzed theoretically VG, and concluded that it is independent to the gradient and capture higher order partial derivatives.

In general, VG is defined as follow

$$\phi^{SG}(f, x) := Var(\phi(f, x + z_j)) \tag{2.17}$$

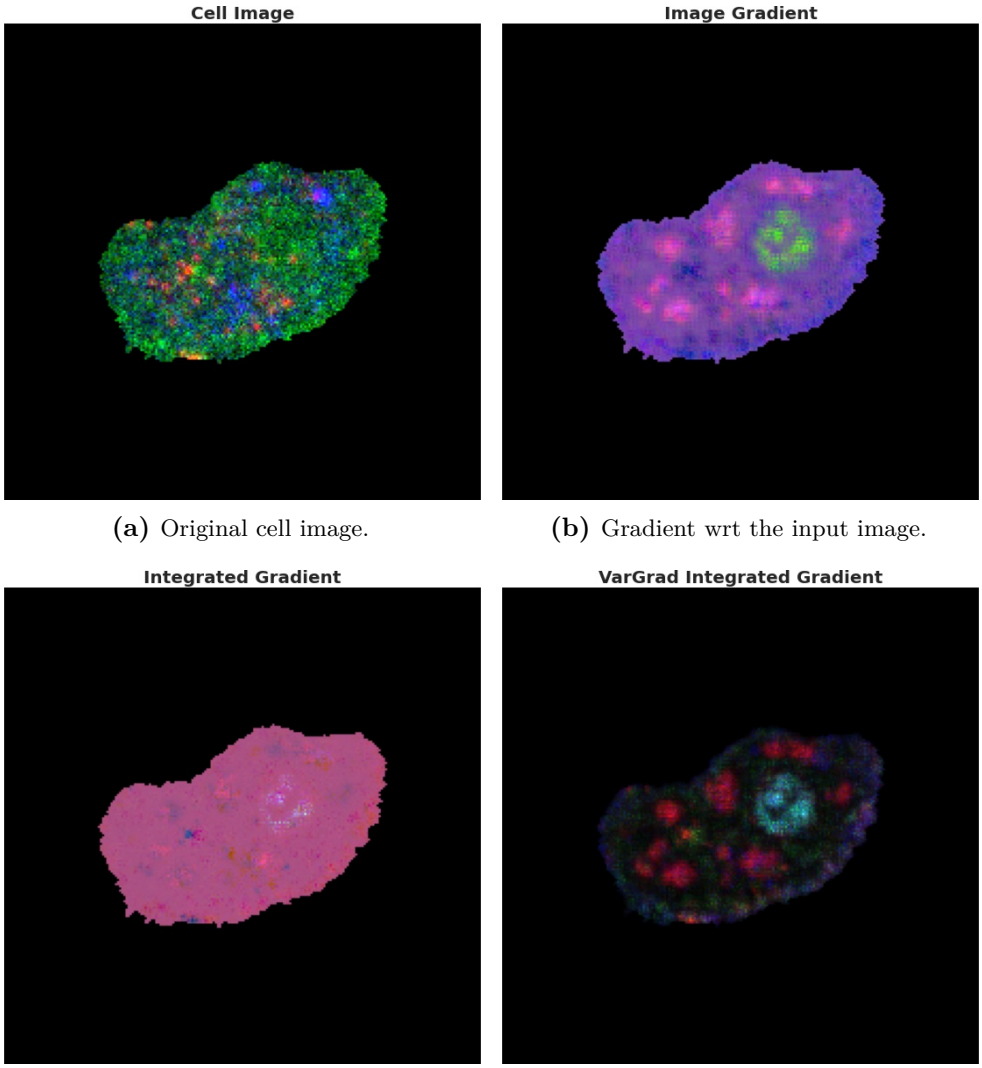
where $x \in \mathbb{R}^{d \times d \times c}$ is the input image, $f : \mathbb{R}^{d \times d \times c} \to \mathbb{R}$ a model, ϕ an attribution method to get preliminary score maps and $z_j \sim \mathcal{N}(0, \sigma^2)$, with $j \in \{1, \ldots, n\}$, are i.i.d. noise images of same shape as the input image.

Since we use IG to get preliminary score maps, in our case VG (in the subsequent defined as VarGrad Integrated Gradient (VGIG)) looks as follow

$$\phi^{SG}(f, x) := Var(\phi^{IG}(f, x + z_j, x')) \tag{2.18}$$

where $x' \in \mathbb{R}^{d \times d \times c}$ is a given baseline needed to compute the IG score maps.

Figures 2.18c and 2.18d show a comparison between IG and VGIG score maps. One can see that VGIG produces less noisy score maps than vanilla IG.



(a) Original cell image.

(b) Gradient wrt the input image.

(c) Integrated Gradient.

(d) VarGrad with Integrated Gradients.

Figure 2.18: Comparison between a cell image and the different attribution methods. All the figures show the same 3 channels taken from a cell image. a) cell image, i.e. no attribution method. b) score map using only the gradient of the model with respect to the input image. c) Integrated Gradient score map. d) VarGrad Integrated Gradient score map.

Chapter 3

Dataset

We can interpret Transcription Rate (TR) as the amount of new RNA molecules inside a cell nucleus in a given period of time. By means of a fluorescent marker, it is possible to identify these new RNA molecules and thus approximate TR. But, what about the morphology of other molecules and organelles within the cell nucleus? The distribution, shape and location of molecules, proteins and organelles within the nucleus could potentially encode relevant information for cellular expression. This has been the main motivation for this work. By means of a Convolutional Neural Network (CNN), we seek to predict TR base mainly in spacial information encoded on images of cell nucleus.

In this section we introduce the process used to generate the data for this work, the multiplexed protein map (MPM) protocol. In addition to this, we introduce the preprocessing and data augmentation techniques used. These techniques aim to improve the model's training performance, prevent overfitting and remove non-relevant information from the images. With this, we seek to encourage the model to base its prediction mainly on the spatial information encoded in the images of cell nucleus.

3.1 Multiplexed protein maps

The amount of protein or messenger RNA (mRNA) inside a cell may not be enough to fully describe cellular function. Accordingly to Buxbaum et al. [BHS14] and Korolchuk et al. [Kor+11], cellular function can heavily depends on the specific intracellular location and interaction with other molecules and intracellular structures. Therefore, cellular expression is determined by the functional state, abundance, morphology, and turnover of its intracellular organelles and cytoskeletal structures. This means that having the ability to look at the concentration and distribution of different molecules within a cell, is an important technological achievement that can significantly leverage scientific discoveries in biomedicine. This is exactly what multiplexed protein map (MPM) allows us to do ([GHP18]). MPM are protein readouts from cell cultures, that simultaneously captures different properties of the cell, like its shape, cycle state, detailed morphology of organelles, nuclear subcompartments, etc. It also captures highly multiplexed subcellular protein maps, which can be used to identify functionally relevant single-cell states, like Transcription Rate (TR). These maps can also identify new cellular

states and allow quantitative comparisons of intracellular organization between single cells in different cell cycle states, microenvironments, and drug treatments [GHP18].

So, let us explain more in deept what are these MPM. Accordingly to Gabriele Gut et al. [GHP18], MPM is a nondegrading protocol that allows to capture efficiently thousands of single cell multichannel images, where each channel contains the distribution and concentration of a protein of interest inside each cell. To achieve this, the protocol is made up of different steps that will be briefly explained here.

Iterative indirect immunofluorescence imaging

The MPM protocol starts with a process called iterative indirect immunofluorescence imaging (4i) developed by the same group. The 4i is a complete protocol by itself, and it allows to capture the concentration and distribution of individual proteins in thousands of different cells in a tissue[1]. Before applying the 4i protocol, the *plate* where the cell culture is must be divided into squared sections called *wells*. Then, the 4i protocol is applied over each well and photographed in sections called *sites*.

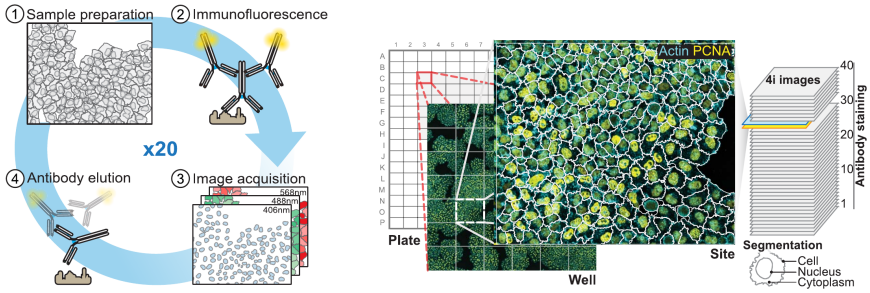
Roughly speaking, 4i works as follow

1. The selected well is prepared for the staining-elution process.
2. The well is saturated with a liquid containing *antibodies*[2] stained with a fluorescent ink (Indirect immunofluorescence (IF)), which binds to a target protein.
3. The well is exposed to a high-energy light and photographed using a light microscopy (which produces a single channel image).
4. The antibodies inside the tissue are washed-out using a chemical elution substrate.
5. Steps 2 to 4 are repeated 20 times to get 20 images of the same protein.
6. To improve the protein readouts, the 20 single channel images are projected into one by *maximum intensity projection*.

Figure 3.1a illustrates the steps of the 4i protocol that capture the saturation and distribution of a specific protein. Keep in mind that even though the 4i protocol captures sever images of the tissue, it returns an uni-channel image (step 6). Figure 3.1b shows the 4i protocol applied 40 times with different IF and over a 384-well plate, which captures the concentration and distribution of 40 different specific proteins.

[1] The tissues were made from cell cultures using the *HeLa Kyoto 184A1* cell line. HeLa is the oldest and most commonly used immortal human cell line in scientific research. The story behind it is quite interesting, so it's worth checking out.

[2] An antibody is a Y-shaped protein that can recognize and bind to a unique molecule (its antigen, e.g. another protein).



(a) Iterative indirect immunofluorescence imaging (4i) protocol.

(b) 4i protocol applied over a specific well of a plate and for 40 different IF.

Figure 3.1: Schematic representation of the 4i protocol for a single well and for 40 different fluorescent antibodies. Figure b also shows the image analysis to identify single cells and its components (nucleus and cytoplasm). Images source [GHP18].

By the time [GHP18] was published, the 4i protocol was able to capture cell culture images with up to 40 channels without degrading the tissue, which is why MPM is called a *nondegrading* protocol.

Multiplexed single cell analysis

Once the multichannel images were generated using the 4i protocol, a series of image preprocessing and image analysis methods ([Car+06] and [Sni+12]) are applied to generate segmentation masks to identify individual cells, as well as their cytoplasm and nucleus. Figure 3.1b shows this segmentation at a cellular level, while figure 3.2 shows it also at a subcellular level. In both cases the boundaries are marked with a white contour. This single cell analysis is also used to identify cells that do not satisfy certain quality controls (like cells in the border of the image or in mitosis stage). However, this will be addressed in detail on section 3.2.

Cell cycle phase classification: G_1, S, G_2 and M phase

The MPM protocol is not only capable to capture the concentration and distribution of molecules inside thousands of cells. It can also identify the phase each cell is in, which is tightly related with the abundances and distribution of molecules inside a cell [GHP18].

Roughly speaking, cell cycle phase was determined by means of a Support Vector Machine (SVM) classifier and k-means clustering. First, a SVM classifier is trained to identify M phase cells based on the nuclear information in one of the image channels

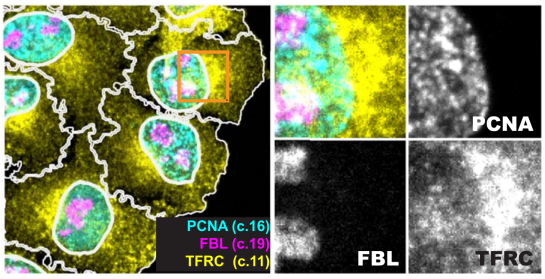


Figure 3.2: Visualization of the subcellular segmentation of a 4i protocol for 18 IF stains. The image was created by combining the readouts of 3 of this IF stains: PCNA (cyan), FBL (magenta) and TFRC (yellow). The number next to each staining label indicates their corresponding 4i acquisition cycle (4i protocol step 5). The orange rectangle and the tile at its right shows a section of the nucleus and cytoplasm of a single cell. The other 3 tiles shows the 4i readout of each of the 3 proteins. Images source [GHP18].

($DAPI$[3]). Then, based on the nuclear information of channel $PCNA$, a second SVM classifier is trained to identify cells in phase S. Finally, cells in phase G_1 and G_2 are classified using a k-means algorithm, using the pixel intensity profiles of the DAPI channels excluding the cells in S and M phase. A more detailed explanation of the cell cycle classification process can be found on the dataset paper [GHP18].

Multiplexed protein maps experimental setup

Although the data used for this work were generated largely following the MPM protocol, they are not the same as those shown in [GHP18]. We warmly thank **Scott Berry** from Pelkmans Lab (at the University of Zurich), for providing the data for this work.

As we already mentioned, the MPM protocol is capable of capturing up to 40 different proteins and molecules within a cell nucleus using fluorescent markers. The MPM data provided for this work contains 38 channels, .i.e. readouts of 38 different proteins and molecules. Table A.3 in appendix A.2, shows the marker used for each channel. Table B.1 (appendix B.1), shows an explanation of some of these markers.

As we mentioned on section 1.3, the MPM protocol also include the use of pharmacological and metabolic perturbations to some sections of the cell culture. However, this work focused on cells without those perturbations. This means that only cells marked as *normal* (no perturbed cells) and $DMSO$[4] (control cells) were used.

As mentioned in this section, the 4i protocol is applied to cell cultures, which are

[3] A brief description of this marker can be found on section B.1.

[4] Dimethyl sulfoxide, or DMSO, is an organic compound used to dissolve test compounds in in drug discovery and design [Cus+20].

divided into rectangular sections called *wells*. The results shown in chapter 5 were obtained using the cells belonging to wells *J16, I14, J10, I09, I11* and *J12*, which also correspond to the unperturbed and control cells (*normal* and *DSMO*). These wells provided a total of 3,703 cells.

Recall that one of the objectives of this work is to fit a model capable of estimating the TR of a cell, given a snapshot of it at a specific moment of time. However, the TR is not provided directly with the MPM data. Instead, for each cell its TR is calculated by averaging the measured pixels corresponding to the marker *EU* (in or case channel 35, see tables A.3 and B.1). Marker `EU` contains nuclear readouts of nascent RNA molecules (pre-messenger RNA (pre-mRNA)) in a given period of time. For the data provided, this time period is 30 minutes, and is the same for all the cells.

3.2 Data preprocessing

The data preprocessing consist of 4 main steps

1. The raw data processing, where raw files are converted into images.
2. The quality control, where cells that are not useful for analysis are discarded.
3. The creation of the dataset, where data is spitted into *Train, validation* and *Test* sets and stored in a way that can be used for model training efficiently.
4. The image preprocessing, where the images are prepared before training the model (clipping and standardization).

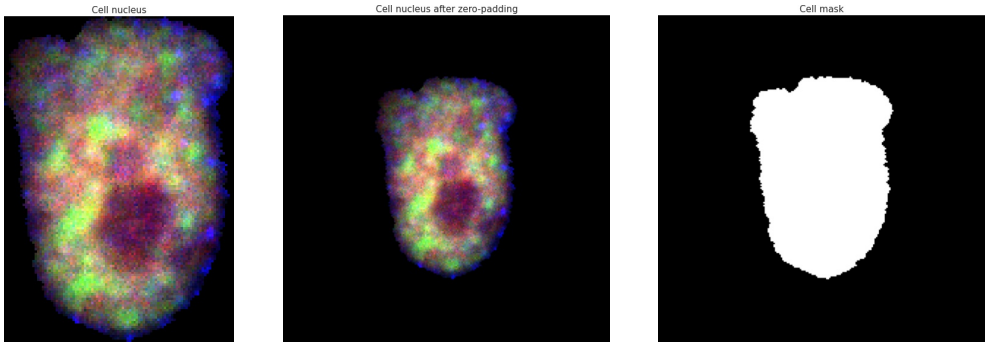
In this section we explain these 4 steps. However, the implementation is discussed in the sections A.1 (for steps 1 and 2) and A.2 (for steps 3 and 4).

3.2.1 Raw data processing

As we mentioned in section 3.1, the multiplexed protein map (MPM) protocol is applied over section of cell cultures called *wells*. The MPM protocol will return several files for each well, containing the nuclear protein readouts of single cells, information from the subsequent analysis made to the intensities of the protein readouts, as well as information about the MPM protocol experimental setup. We do not go into details about this files and how to transform them into multichannel images of single cell nucleus. However, a brief explanation of this can be found in the appendix A.1. Appendix A.1 also show how to run the Python script that transforms the raw data into images, along with an explanation of the required parameters.

The Python script introduced on appendix A.1 extract the protein readouts from the raw data files, and use them to build multichannel images containing the nucleus of

a single cell (see figure 3.3a). This means that during the reconstruction of the images, it is necessary to add black pixels (zeros) in the places where no measures were taken (like in the low corner of figure 3.3a). However, as we saw on section 2.2.2, in order to train a Convolutional Neural Network (CNN) model, all the cell images need to have a fixed size, which is denoted as I_s. For this reason, after the image is reconstructed, it is necessary to add zeros to the images borders (zero-padding) in order to make it squared and of a fixed size (see figure 3.3b). Finally, for each single cell nucleus, a *cell mask* is created to keep track of the measured and non-measured pixels (see figure 3.3c). As we can see in figure 3.3, the cell nucleus is always located in the center of the image.



(a) Single cell nucleus.

(b) Single cell nucleus with zero-padding.

(c) Single cell nucleus mask.

Figure 3.3: Figure a shows channels 10, 11 and 15 of the nucleus of a single cell multichannel image reconstructed form the raw data. Figure b shows image a after adding zero to the borders (zero-padding) to make it of size 224 by 224 pixels. Figure c shows the cell mask, i.e. measured pixels (in white) during the MPM protocol.

The raw data processing script saves in a specified directory files containing 3 compressed NumPy arrays; 1) the multichannel image (figure 3.3b), a 3D array contains the protein readouts of the nucleus of a single cell 2) the cell mask (figure 3.3c), a 2D array that indicates the measured pixels by the MPM protocol (ones on the measured x and y coordinates and zeros otherwise) and 3) the channels average, a 1D array containing the average of the measured pixels per channel/protein. Each file is named using the unique id assigned to each single cell nucleus (`mapobject_id_cell`). The script also returns a `csv` file[5] containing the metadata of each single cell from every processed well (one row per cell and one column per cell feature). Table 3.1 shows the metadata columns that were relevant for this work.

[5]This `csv` file can be easily opened as a *Pandas DataFrame*. For more information, please refer to the official documentation.

Column name	Description
`mapobject_id_cell`	ID to uniquely identify each cell among all wells
`mapobject_id`	ID to uniquely identify each cell on its well
`is_border_cell`	Binary flag, 1 if the cell is on the plate, well or site border; 0 if not
`cell_cycle`	String, `G1` if cell is in G_1 phase, `S` if cell is in synthesis phase, `G2` if cell is in G_2 phase. If `NaN`, then the cell is in mitosis phase
`is_polynuclei_184A1`	Binary flag for *184A1* cells, 1 if the cell was identified to have more than one nucleus (i.e. it is in mitosis phase); 0 if not
`is_polynuclei_HeLa`	Binary flag for *HeLa* cells, 1 if the cell was identified to have more than one nucleus (i.e. it is in mitosis phase); 0 if not
`perturbation`	String indicating the pharmacological/metabolic perturbation

Table 3.1: Relevant metadata columns.

3.2.2 Quality Control

During the transformation from raw data into images, cells that does not pass a quality control are discriminated. This quality control consist on avoiding cells that holds at least one of the following conditions

1. The cell is in mitotic phase (i.e. on metadata, either `is_polynuclei_HeLa` or `is_polynuclei_184A1` is equal to 1 or `cell_cycle` is `NaN`).
2. The cell is in the border of the plate, well or site (i.e. on metadata, `is_border_cell` is equal to 1).

The quality control is performed by the same script that transforms the raw data into multichannel images. Its implementation and execution, as well as an explanation of the required parameters, can be found on appendix A.1.

3.2.3 Dataset creation

After the raw data from all wells were processed, and mitotic and/or border cells were eliminated (quality control), we are able to build a dataset[6] that can be used efficiently

[6]For this work we decided to use (and build) a custom TensorFlow Dataset (TFDS), which is a subclass of `tensorflow_datasets.core.DatasetBuilder` and allows to create a pipeline that can easily feed data into a machine learning model built using TensorFlow. For more information, please refer to the official documentation.

to train models. We will not explain here how to create this dataset. However, a brief explanation of this can be found in the appendix A.2. Appendix A.2 also show how to run the Python script that builds this dataset, along with an explanation of the required parameters.

Even though this script can bu used to build a dataset containing all available single cell images, for this work we created a dataset containing cells without pharmacological or metabolic perturbations (i.e. cells such that in the metadata `perturbation` is ether equal to *normal* or *DMSO*). Further more, during the creation of the dataset, it is possible to filter the image channels and select the target value from the channels average vector (which is stored along with each single cell image). In this case we kept all the input channels[7], except for the channel used to calculate the target value. This means that channel 35 was excluded (`00_EU`[8]), and entrance 35 from the channel average vector (interpreted as Transcription Rate (TR)) was selected as target value.

Last but not least, for each cell, its mask is added at the end as an extra channel to keep track of the measured pixels. The reason why the cell mask is stored as a channel, is because it will be needed by other process latter in the pipeline (some of the data augmentation techniques, see section 3.3). However, this (and other channels) are removed before the image is used to feed the model (during and after the training process, see section 4.2).

Table A.3 (on appendix A.2) shows the image channels in the TensorFlow Dataset (TFDS), including the name (column *Channel name*) and identifier of each immunofluorescence markers (column *Marker identifier*). Table A.3 also shows the ids corresponding to the markers in the raw data (column *Raw data id*) and in the TFDS (column *TFDS id*). *NA* means that the channel is not used/available either on the raw data or the TFDS.

Set	Number of elements	Percentage
Train	2962	80%
Validation	371	10%
Test	370	10%
Total	3703	100%

Table 3.2: Distribution of the dataset partitions.

During the creation of the dataset, the images are also spitted into 3 sets, *Train, Validation* and *Test*, using the proportions 80%, 10% and 10% respectively. Table 3.2

[7]The unnecessary/unwanted channels are removed during the model training/evaluation (see section 4.2). The reason why this filtering is not made during the dataset creation, is to make the dataset set more robust (i.e. to avoid the need to create a new dataset each time the input channels of the image changed).

[8]A brief description of this marker can be found on section B.1.

shows the size of these 3 sets.

Set	Cell Cycle	Number of elements	Percentage
Train	G_1	1652	55.77%
	S	864	29.17%
	G_2	446	15.06%
Validation	G_1	205	55.41%
	S	103	27.84%
	G_2	62	16.76%
Test	G_1	213	57.41%
	S	103	27.76%
	G_2	55	14.82%
Total	G_1	2070	55.90%
	S	1070	28.90%
	G_2	563	15.20%

Table 3.3: Distribution of the dataset partitions by cell phase (cell cycle).

Set	Perturbation	Number of elements	Percentage
Train	Normal	2040	68.87%
	DMSO	922	31.13%
Validation	Normal	257	69.46%
	DMSO	113	30.54%
Test	Normal	260	70.08%
	DMSO	111	29.92%
Total	Normal	2557	69.05%
	DMSO	1146	30.95%

Table 3.4: Distribution of the dataset partitions by perturbation.

Since we are dealing with cells in different phases (cell cycles), it is important that the distribution of the 3 phases is kept in the train, validation and test sets. The same must happen with the proportion of cells without pharmacological/metabolic perturbation (*Normal* cells) and control cells (*DMSO* cells). Tables 3.3 and 3.4 show respectively that these proportions are hold across the 3 sets.

3.2.4 Image preprocessing

In this work we use CNNs and images of cell nucleus to predict TR. This means that there are two main features of the images that came into account when the model learns and predicts the TR, the spatial distribution of the elements in the image and

the intensity of the colors. However, this work aims to explain and predict transcription based on the information encoded in the spatial distribution of proteins and organelles within the nucleus. Therefore, the image preprocessing techniques applied here should help mitigate the influence of color during training and prediction, so that the model can focus only on spatial information. For this reason, two preprocessing techniques are applied to each cell image, clipping and standardization. The clipping, as well as the standardization, are performed during the construction of the TFDS, which can be consulted in appendix A.2.

Clipping

The idea of clipping is to avoid extreme outliers to influence or leverage the model parameters during training. Figure 3.4 gives an example of this. The blue line shows a model fitted including the outliers (the two dots on the right upper corner), while the orange line a model fitted without them.

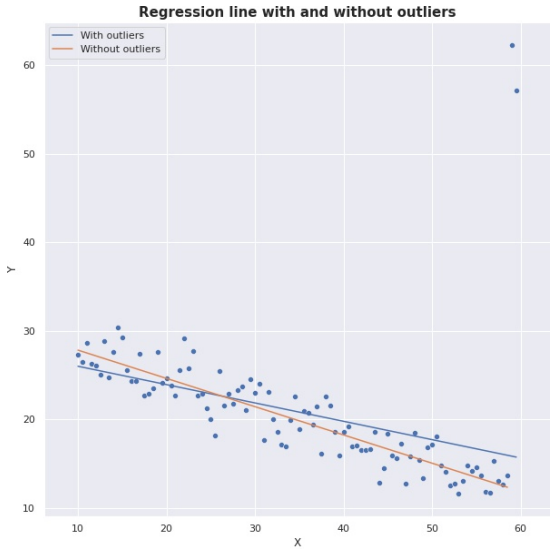


Figure 3.4: Comparison between two linear regression models, fitted with (blue line) and without (orange line) outliers.

To prevent high intense pixels to influence the model, we truncate/limit the value of pixels that are above a certain threshold. This threshold is different for each image channel and is determined using the cell images belonging to the training set. For each channel, the train images are loaded and the threshold is set as the 98% percentile of the measured pixel intensities belonging to the channel. Then, using this threshold vector (one entrance per channel) all the images in the dataset (train, validation and test) are clipped. This is done before the data standardization. Finally, the clipping

parameter (threshold) of each channel is stored in a metadata file, provided along with the TFDS. Figures 3.5a and 3.5b show the pixel intensity distribution of channel HDAC3 before and after clipping respectively.

Standardization

As we mentioned at the beginning of this section, to predict cell TR we seek the model to rely on spatial information, rather than the intensity of the pixels. Therefore, to reduce pixel intensity influence, we apply per-channel standardization, which is just a shift and rescaling (a linear transformation) of the original data. Standardization is also called *Z-score*, since the data is transformed using the mean μ and standard deviation σ (normal distribution parameters) of a sample, as a shift and rescaling parameters respectively. As it is done in clipping, the standardization parameters are different for each channel and are computed using the images belonging to the training set. For all the measured pixels intensities in the TFDS (i.e. for train, validation and test sets), the standardization of pixel i belonging to channel c (i.e. $z_{i,c}$), is done as follow

$$z_{i,c} = \frac{x_{i,c} - \mu_c}{\sigma_c} \tag{3.1}$$

where $x_{i,c}$ is the corresponding readout i from channel c, and μ_c, σ_c are the mean and standard deviation (respectively) of channel c computed using the training images.

The standardization centers the measured pixels of each channel around 0 (see figures 3.5b and 3.5c), reducing the color correlation between channels, which also reduce pixel intensity influence over the model.

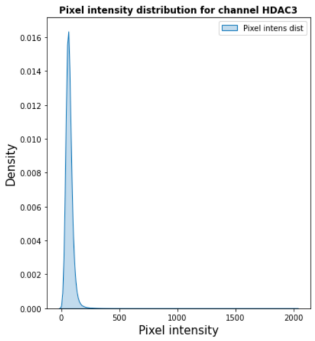
Figure 3.6 shows 3 different cell nucleus sampled from the resulting TFDS. Each nucleus is in a different cell phase (G_1, S and G_2 respectively), and shows a different group of 3 markers (channels).

3.3 Data augmentation

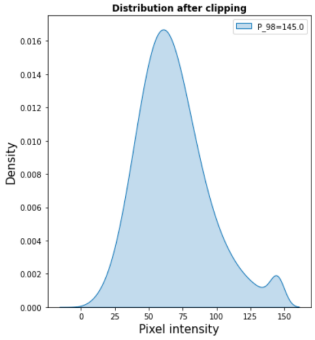
Data augmentation techniques are widely used to improve the results of Convolutional Neural Network (CNN) by reducing overfitting ([KSH17], [SSP+03]). This techniques improve generalization by creating new date from the existing one.

However, data augmentation techniques can not only help us to prevent overfitting, they can also be used to remove characteristics of the data that are not of interest to us. In our case, we want the model to rely on spatial information, rather than pixel intensity (color). Therefore, we implement the following data augmentation techniques, which help us to achieve this

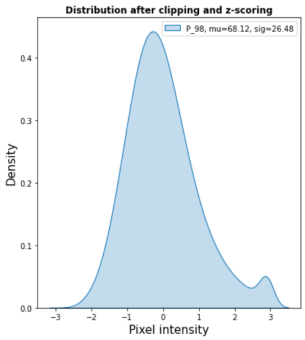
- To remove non-relevant data features



(a) Pixel intensity distribution.



(b) Pixel intensity distribution after clipping.



(c) Pixel intensity distribution after clipping and standardization.

Figure 3.5: Intensity distribution of measured pixels for channel HDAC3. The channel readouts were taken from the training set. Figure a) shows the distribution without any modification. Figure b) shows the distribution after applying 98% percentile clipping, while figure c) shows the distribution after applying same clipping and standardization.

  - Color shifting
  - Image zoom in/out

- To improve model generalization
  - Horizontal flipping
  - 90 degree rotations

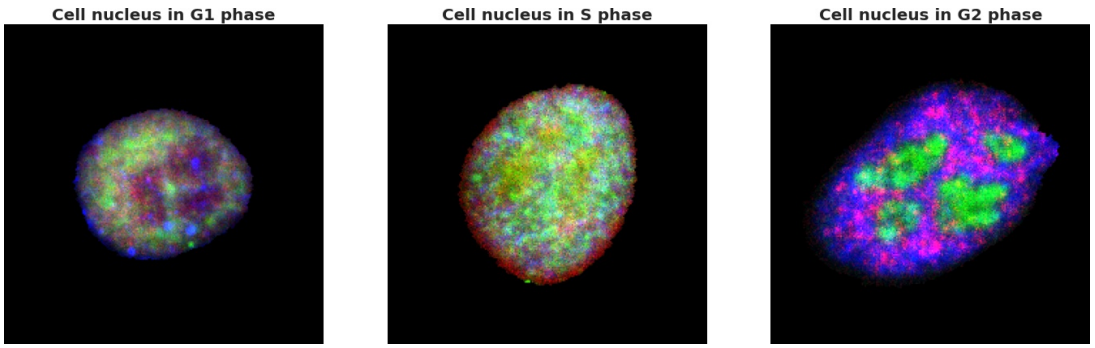


Figure 3.6: Cell nucleus in phases G_1, S and G_2 respectively. Each nucleus shows a different group of 3 markers.

This techniques are applied during training time[9] and in the same order as shown in the list above.

Nevertheless, data augmentation techniques are not only limited to the training set. As we shown above, we can divide them into two groups; 1) to remove non-relevant data features and 2) to improve model generalization. So it makes sense to apply the second group to the validation set, so we can get a better idea of how well the model is generalizing during training[10]. This means that the size of the validation set is increased by a factor of 8, by rotating each element 0, 90, 180, 270 degrees and applying (or not) a horizontal flipping. Note that the transformations performed on the validation set do not introduce any randomness to it.

3.3.1 Color shifting

The data preprocessing techniques introduced on section 3.2 (clipping and standardization), helped to reduce the influence that the intensity of the pixels, as well as the correlation between channels, have on the prediction of the model. By doing this, we encourage the model to rely more on the spatial information encoded in the images.

However, we can go a little further by shifting the pixel intensities by a random number, which would reduce even more the influence of color on the model prediction. If we sample a different random number for each channel, then the correlation between channels is also reduced. We must be careful here, since during raw data processing (see section 3.2.1) zero pixels were added to reconstruct the images. Therefore, if we add a different random number to each channel, then the non-relevant (unmeasured) pixels will have different values. Fortunately, during the creation of the TensorFlow Dataset (TFDS), for each cell we included its *cell mask* to its image as another channel (the last one). Therefore, we can use this information to randomly shift only the measured pixels. Mathematically, this means that for the channel c its $i-th$ measured pixel $x_{i,c}$, the shifted pixel $x'_{i,c}$ is defined as

$$x'_{i,c} = x_{i,c} + \eta_c \tag{3.2}$$

where $\eta_c \sim U(-a, a)$, with $c \in \{0, \ldots, C\}$, are i.i.d. random variables.

Figures 3.10a and 3.10b show a cell nucleus image before and after applying per-channel random color shifting respectively.

[9] This means that instead of generating new data and then adding it to the dataset before training, new data is generated *on the fly* during training, by applying random predefined transformations to the existing data.

[10] As always in statistics, more data equals more accurate approximations.

3.3.2 Image zoom-in/out

Size is another characteristic of the cell nucleus that could influence the output of the model. Figure 3.7 shows three cell nucleus with different sizes. However, as we already mentioned, we seek the model to predict Transcription Rate (TR) based on the distribution of organelles and proteins inside the nucleus (spatial information). For this reason, we randomly zoom-in/out the image to either increase or decrease (upsample or dawnsample respectively) the size of the cell nucleus inside it. This zoom is always applied over the center of the image. After that, the image either must be cropped in the center, or add zeros in the borders (padding), so the size of the zoomed image match the original size, i.e. I_s. Since this randomize the cell nucleus size, the model can not longer rely on it to make a prediction. Figures 3.10a and 3.10c show an example of this.

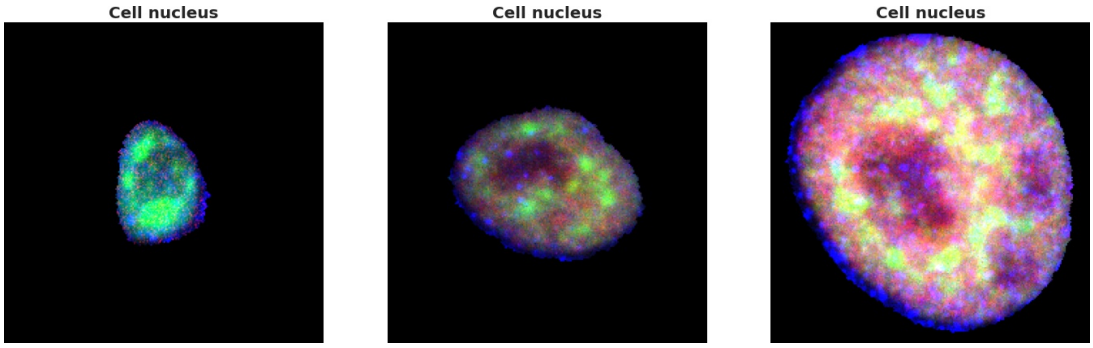


Figure 3.7: Cell nucleus with different sizes.

However, there are two things to have in mind when the size of the cell nucleus is changed, the maximum zoom-in (to avoid cutting the cell nucleus borders) and the cell nucleus size distribution.

To avoid zooming-in over a cell nucleus image too much and cut its edges, we need to determine the maximum zoom-in ratio U_{max} (which is different for every image of a cell nucleus). This can be computed as follow

$$\begin{aligned} U_{max} &:= 1 - S_{ratio} \\ &:= 1 - \frac{2d_{min}}{I_s} \end{aligned} \tag{3.3}$$

where $d_{min} := min\{a, b, c, d\}$ is the minimum distance between the cell nucleus and the image borders. Figure 3.8 illustrates this distances.

Intuitively, $S_{ratio} := 2d_{min}/I_s$ (cell nucleus size ratio) denotes the proportion that the cell nucleus is occupying in the image (transberzally).

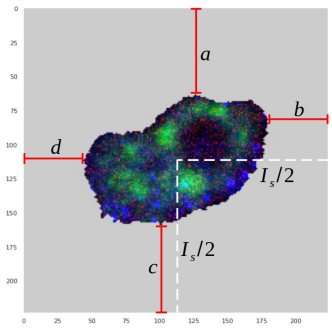


Figure 3.8: Distances needed to determine the cell size ratio. The red lines show the distance between the measured pixels of the cell nucleus (border pixels) to the 4 edges of the cell image. The white dashed lines indicates the center of the image.

The last thing that has to be considered, is the distribution of the cell nucleus sizes. Since we are randomizing them, the distribution of the new sizes must be similar to the original distribution. Fortunately, during the raw data processing (see section 3.2), the cell nucleus size ratio S_{ratio} of each cell was computed and saved in the metadata. Figure 3.9 shows the distribution of S_{ratio}. During model training, the zoom-in/out proportion is sampled considering this distribution.

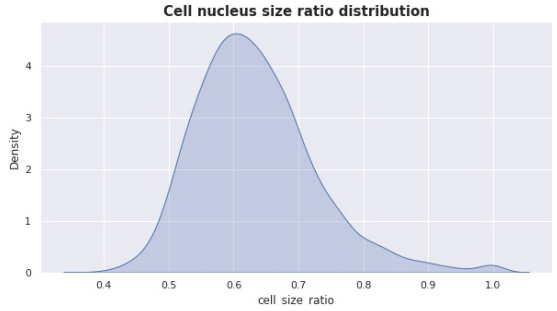


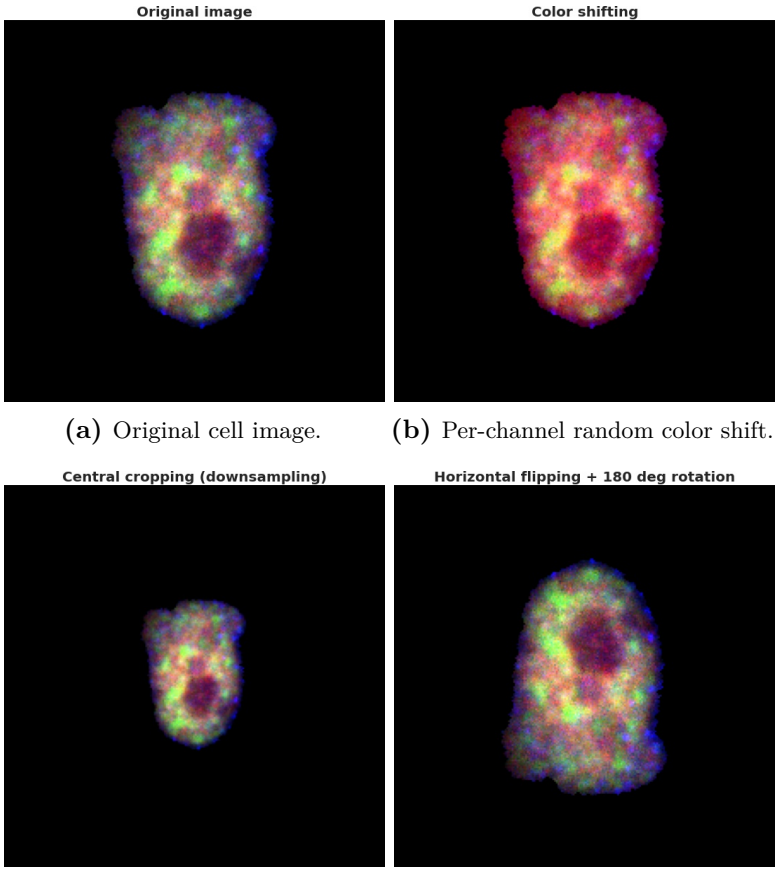
Figure 3.9: Cell nucleus size ratio S_{ratio} distribution.

3.3.3 Horizontal flips and 90 degree rotations

Since there is no sense of orientation in a cell (there is no top, bottom, left, or right), flips and rotations will not change the distribution of the data at all. For this reason, we can use these transformations to simply increase the amount of data and alleviate overfitting.

For this work we used random horizontal flips and $k \times 90$ (degree) rotations, for $k \in 0, 1, 2, 3$. As we already mentioned, these transformations are applied in both the

training and the validation sets.[11]. Figures 3.10a and 3.10d shows an example of a cell nucleus image after being flipped and rotated 180 degrees.



(a) Original cell image.

(b) Per-channel random color shift.

(c) Central cropping (downsampling).

(d) Horizontal flipping + 180 degree rotation.

Figure 3.10: Data augmentation techniques. Figure a) shows channels 10, 11 and 15 of a multichannel image without augmentation techniques. Figure b) shows image a) after applying per-channel random color shifting. Figure c) shows image a) after applying central cropping (in this case, downsampling). Figure d) shows image a) after applying horizontal flipping and 180 degree rotation (counter-clockwise).

[11]The only difference is that for the validation set, the flips and rotations are applied deterministically, while for the training set they are applied randomly.

3.4 Discussion

To identify nascent RNA inside a cell nucleus, the multiplexed protein map (MPM) protocol use the 5-ethynyl uridine (EU) marker (which is then interpreted as Transcription Rate (TR)). However, it has been observed that this marker also binds to DNA molecules after some incubation time [JS08], [Bao+18]. As future work, another MPM dataset could be analyzed, either with a shorter or longer incubation time for the UE marker. Then, it would be interesting to validate if the results obtained with both datasets are consistent.

Besides the preprocessing techniques introduced in section 3.2 (clipping and standardization), the following approaches were also tried

- Linear scaling using the 98% percentile with and without clipping.
- Mean extraction and linear scaling using the 98% percentile with clipping (like standardization, but with the 98% percentile instead of the standard deviation).
- 49% percentile extraction and linear scaling using the 98% percentile (no clipping).

This approaches were tried at a per-channel level. However, clipping plus standardization where the prprocessing techniques that showed the best performance. Since we seek the model to predict TR base on spacial information, rather than pixel intensity/color, good performance means low Mean Absolute Error (MAE) for the Convolutional Neural Network (CNN) models, but high MAE for the linear model (since the linear model is unable to use the spatial information). This indicates that the spatial information encoded in the images of the data set has more influence on the prediction of the model than the information encoded in the colors.

Another aspect of the dataset that is worth to mention, is that more than half cells are in phase G_1 (see table 3.3), while cells in S phase are less than 30% and around 15% for G_2 cells. This causes the model to focus more on correctly predicting the TR of G_1 cells, than for cells in the other two phases. This happens because G_1 cells have more influence on the minimization of the objective function, since it is more likely that the model is fed with G_1 cells during training.

As it is shown on figure 3.11, the TR of G_1 cells is significantly lower than the TR of S and G_2 cells. This, and that cells in different phases are not in the same proportion in the dataset, could cause the model to make a biased prediction when it is fed with a S or G_2 cell. Two possible solutions to this problem are, either to add more cells in phases S and G_2 to the dataset, or to sample with replacement over the available cells, so the proportion of cells in the three different phases is the same in the dataset. Another possible solution would be to make a weight loss function based on the proportions of the cell phases, such that every phase has the same influence on it during training.

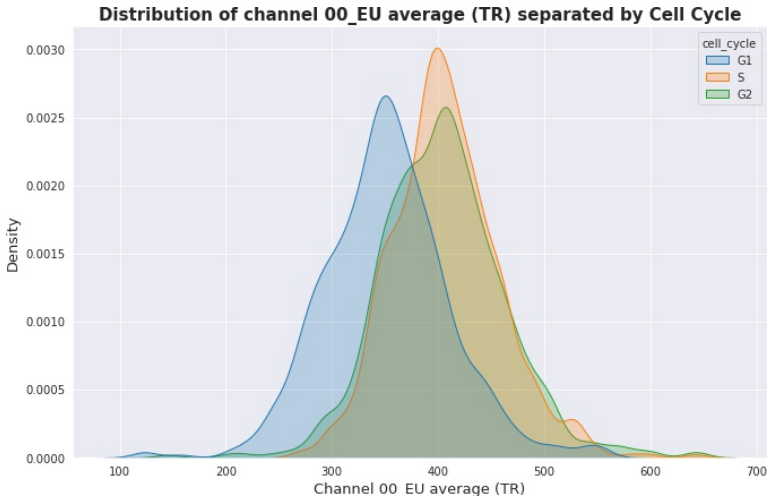


Figure 3.11: TR distribution separated by cell phase.

In [Smi+17], Smilkov et al. mention that the addition of random noise to the images during model training, can improve the quality of the score maps (less noisy score maps). For this reason, beside the data augmentation techniques introduced in this section, the addition of random noise was also implemented and tried. However, in practice this did not show any apparent improvement.

Chapter 4

Methodology

This chapter provides the experimental setups for each element of the workflow shown in figure 4.1. This includes the raw data processing (selected raw data and the quality control), the TensorFlow Dataset (TFDS) parameters, the data augmentation techniques (as well as their hyperparameters), the used models (architecture, loss function, optimizer and hyperparameters), the metrics used to evaluate the performance of the models and the hyperparameters corresponding to the interpretability methods.

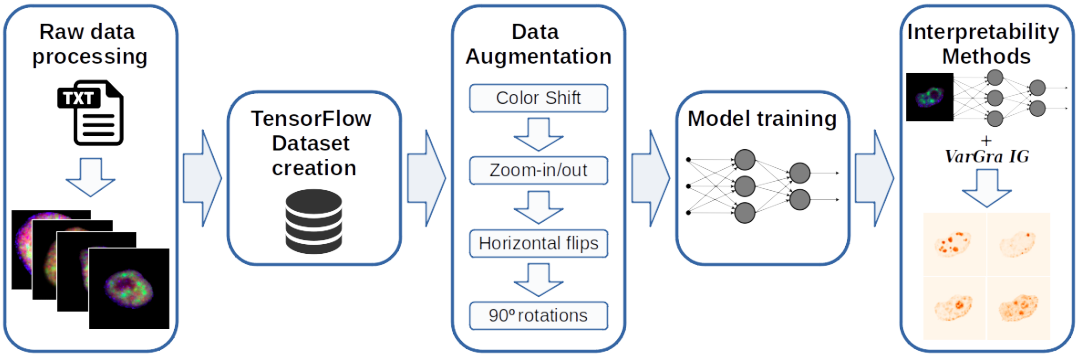


Figure 4.1: Workflow.

4.1 Dataset Setup

In this section we specify all the hyperparameters needed to execute the process explained on chapter 3. This contemplates the raw data processing, the quality control, the TensorFlow Dataset (TFDS) creation, the image preprocessing, as well as data augmentation.

4.1.1 Data preprocessing

As we explained in section 3.2, the data preprocessing consist of 4 main steps; 1) the raw data processing, 2) the quality control, 3) the creation of the dataset and 4) the

image preprocessing.

A complementary explanation of the data preprocessing parameters, as well as implementation references, can be found in the appendices A.1 and A.2.

Raw data processing

As we explained in section 3.2.3, to build the TFDS it is necessary to specify the perturbations that will be included in the dataset. For this reason, all the available wells were processed and transformed into images. This included wells exposed to pharmacological and metabolic perturbations, control wells and unperturbed wells. This allows the user to easily create new datasets without having to run the raw data processing first. Table 4.1 shows the processed wells separated by perturbation.

Perturbation type	Perturbation name	Well names
pharmacological/metabolic	CX5461	I18, J22, J09
	AZD4573	I13, J21, J14, I17, J18
	meayamycin	I12, I20
	triptolide	I10, J15
	TSA	J20, I16, J13
control	DMSO	J16, I14
unperturbed	normal	J10, I09, I11, J18, J12

Table 4.1: Well names divided by perturbation name and type.

Another hyperparameter that needs to be specified during the raw data processing, is the size of the output images I_s. This size applies to both, the width and height of the image (square images). Since some prebuilt architectures use a standard image size of 224 by 224, we define I_s as 224.

Quality control

As it is mentioned in section 3.2.2, the quality control is meant to exclude cells with undesirable features. In our case we discriminate mitotic and border cells. The information used by the quality control is contained in the metadata of each well. Table 4.2 shows the metadata columns and the discriminated values. If a cell has any of these values, then it is excluded.

Dataset creation and image preprocessing

As it is explained in section 3.2.3, in this work we decided to use a custom TFDS. Table 4.3 shows the parameters used to build the dataset employed in this work, together with the image preprocessing parameters.

Feature	Metadata column name	Discriminated value
Cell in mitosis phace	`is_polynuclei_HeLa`	1
	`is_polynuclei_184A1`	1
	`cell_cycle`	`NaN`
Border cell	`is_border_cell`	1

Table 4.2: Discrimination characteristics for quality control.

Parameter	Description
Perturbations to be included in the dataset	*normal* and *DMSO*
Cell phases to be included in the dataset	G_1, S, G_2
Training set split fraction	0.8
Validation set split fraction	0.1
Seed	123 (for reproducibility of the train, val and test split)
Percentile	98 (for clipping / linear scaling / standardization)
Clipping flag	1
Mean extraction flag	0
Linear scaling flag	0
Standardization (z-score) flag	1
Model input channels	All of them except for channel `00_EU` (see table A.3)
Channel used to compute target variable (output channel)	`00_EU` (channel id 35, see table A.3)

Table 4.3: Parameters used to biuld TFDS and image preprocessing.

The custom TFDS created with the parameters specified in table 4.3 is called `mpp_ds_normal_dmso_z_score`.

The Python script that builds the custom TFDS, also returns a file with the image preprocessing parameters (`channels_df.cvs`) (as this is applied at a per-channel level) and information about the channels (channel name, id, etc.). It also returns another file with the metadata of each cell included in the TFDS (`metadata_df.csv`). These files are stored in the same directory as the TFDS files.

In table 4.3 we also specify the channel used to compute the target variable (ground truth), which is the channel corresponding to the marker *EU* (channel id 35, see tables A.3 and B.1). Recall that this channel contains nuclear readouts of nascent RNA (pre-messenger RNA (pre-mRNA)) in a given period of time. For the data provided, this time period was the same for all the cells (30 minutes) and is specified in the *duration* columns of the metadata. Since channel 35 is used to compute the target variable (ground truth), it is removed from the prediction/input channels.

4.1.2 Data augmentation

In this section we specify the data augmentation techniques (see section 3.3) and its hyperparameters used to train all the models of this work. Recall that the techniques are either aimed to remove non-relevant characteristics of the data (color shifting, central zoom in/out) or to improve model generalization (horizontal flipping, 90 degree rotations). Table 4.4 shows this techniques and its hyperparameters grouped by objective and technique. In practice, the augmentation techniques are applied as shown in table 4.4 from top to bottom.

Objective	Technique	Hyperparameter	Description
Remove non-relevant features	random color shifting	distribution	$U(-3, 3)$
	random central zoom in/out	distribution[1]	$N(\mu = 0.6, \sigma = 0.1)$
Improve generalization	random horizontal flipping	NA	NA
	random 90 degrees rotations	NA	NA

Table 4.4: Parameters used for data augmentation techniques. The NA means that there are no hyperparameters for this technique or that there is no further description.

[1]This distribution is used to sample the *cell nucleus size ratio* S_{ratio} (see section 3.3.2) of each cell. However, the parameters for this distribution (mean and standard deviation) were not provided

Even thought we specify the data augmentation hyperparameters here, in practice these are selected for each model and applied during training. However, all the models showed in this work were trained using the techniques and values shown in table 4.4. A complementary explanation can be found in appendix A.3.

In section 3.3 we mentioned that data augmentation techniques can be applied to both the training set and the validation set. However, we also mentioned that only horizontal flips and 90 degree rotations are applied for the validation set. Furthermore, for the training set these techniques are applied randomly, while for the validation set they are applied deterministically. Therefore, table 4.4 only applies to the training set.

4.2 Models

In this section we introduce the models, and its architecture, used in this work. All the models were implemented in TensorFlow 2.2.0. We also specify all the used hyperparameters. Besides this, the appendix A.3 provides a brief explanation of how to train and evaluate all the models introduced here.

In general, all the models where trained using *ReLU* as activation function for the hidden layers. Also, the identity was used as activation function for the last layer. Table 4.5, shows the other general hyperparameters.

Parameter	Description
Number of epochs	800
Early stopping patience	100
Batch size	64
TFDS name	`mpp_ds_normal_dmso_z_score`
Random seed	123
Input Channels	all channels such that its TFDS id is in $\{0, \cdots, 32\}$ (see appendix A.2, table A.3)

Table 4.5: Hyperparameters used in the training of all the models.

Even though that the number of epochs is specified in table 4.5, if the loss function does not improve (decrease) for more than 100 (i.e. *Early stopping patience*) epochs during training, then the training stops.

Table 4.5 also indicate the input channels to by used by the model as predictors. In section 3.2.3 we mentioned that all the image channels (with the exception of channel `00_EU`) were kept during the creation of the TensorFlow Dataset (TFDS). Moreover,

by us. Instead, they were estimated using the information in column `cell_size_ratio` of the TFDS metadata file. Therefore, the `return_cell_size_ratio` flag must be set to 1 (True) during raw data processing, so this column is created (see section 3.2.1 and appendix A.1).

since the data augmentation techniques are only applied to the measured pixels of the cell images, the cell mask was added to the image as the last channel. For this reason the channel filtering process is made inside the model. This means that after the input layer, the models have a *channel filtering layer*, which basically remove the non-selected channels, by projecting the input image from a space of shape $(bs, 224, 224, 38)$ into a lower one of shape $(bs, 224, 224, 33)$. This is done just by performing a matrix multiplication between the input batch $B \in \mathbb{R}^{bs \times 224 \times 224 \times 38}$ and a projection matrix $P \in \{0, 1\}^{38 \times 33}$ (a zero matrix with ones on the diagonal elements corresponding with to the input channels), i.e. $B_{filterd} = BP$.

All the model in this work were trained using the *Huber* loss function

$$\mathcal{L}_\delta(y, f(x)) = \begin{cases} \frac{1}{2}(y - f(x))^2, & \text{for } |y - f(x)| \leq \delta \\ \delta|y - f(x)| - \frac{1}{2}\delta^2, & \text{otherwise} \end{cases} \tag{4.1}$$

where $\delta > 0$ is the value where the Huber loss function changes from a quadratic to linear. The hyperparameter δ was set to 1 for all the models.

Huber loss function is quadratic when the error $a = |y - f(x)|$ is below the threshold δ (like the Mean Squared Error (MSE)), but linear when it is above it. This makes Huber loss less susceptible to outliers. Figure 4.2 shows a comparison between the Huber (in green) and the MSE (in blue) loss functions.

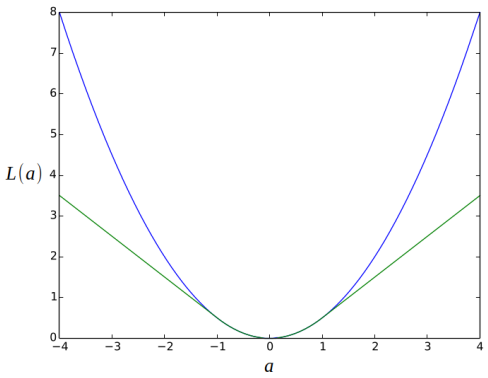


Figure 4.2: Huber (green) and the MSE (blue) loss functions. Image source [Hub].

In section 2.2.1 we mentioned that we choose the Adaptive Moment Estimation (Adam) optimizer to fit the model parameters. With the exception of the *learning rate*, the used parameters were $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e - 07$, which are the default TensorFlow hyperparameters[2]. The learning rate is specified in the section corresponding to each model.

[2]For more information pleases refer to the TensorFlow official documentation.

4.2.1 Linear Model

As we already mentioned, the objective of this work is to explain cell expression using spatial information in multichannel images of cell nucleus. To compare the performance of the Convolutional Neural Networks (CNNs), and have an idea of how much pixel intensity information was still contained in the data, we also fitted a *linear model*

$$y = w_0 + w_1 x_1 + \cdots + w_{33} x_{33} \tag{4.2}$$

where $x_i, \in \mathbb{R}$, for $i \in \{1, \cdots, 33\}$, is the average pixel intensity corresponding to channel i and $w_i, \in \mathbb{R}$, for $i \in \{0, \cdots, 33\}$ are the model coefficients.

The linear model architecture is specified in table 4.6. The rows of the table represent each layer of the model, which are evaluated from top to bottom. This model has only 34 free (learnable) parameters in total and was trained with a learning rate of 0.1.

Layer	Output Shape	Number of parameters
Input	$(bs, 224, 224, 38)$	0
Channel filtering	$(bs, 224, 224, 33)$	0
Global Average Pooling	$(bs, 33)$	0
Dense	$(bs, 1)$	34

Table 4.6: Linear model architecture. The rows represent each layer of the model. The flow of the model is from top to bottom. The *bs* on the *Output Shape* column stands for *Batch size*.

4.2.2 Simple CNN

The *Simple* CNN architecture is specified in table 4.7. The rows of the table represent each layer of the model, which are evaluated from top to bottom. This model has 160,129 free (learnable) parameters in total and it was trained with a learning rate of 0.0005.

Layer	Output Shape	Number of parameters
Input	$(bs, 224, 224, 38)$	0
Channel filtering	$(bs, 224, 224, 33)$	0
Convolution	$(bs, 224, 224, 64)$	19072
Batch Normalization	$(bs, 224, 224, 64)$	256
ReLU	$(bs, 224, 224, 64)$	0
Max Pooling	$(bs, 112, 112, 64)$	0
Convolution	$(bs, 112, 112, 128)$	73856
Batch Normalization	$(bs, 112, 112, 128)$	512
ReLU	$(bs, 112, 112, 128)$	0
Max Pooling	$(bs, 56, 56, 128)$	0
Global Average Pooling	$(bs, 128)$	0
Dense	$(bs, 256)$	33024
Batch Normalization	$(bs, 256)$	1024
ReLU	$(bs, 256)$	0
Dense	$(bs, 128)$	32896
Batch Normalization	$(bs, 128)$	512
ReLU	$(bs, 128)$	0
Dense	$(bs, 1)$	129

Table 4.7: Simple CNN architecture. The rows represent each layer of the model. The flow of the model is from top to bottom. The *bs* on the *Output Shape* column stands for *Batch size*.

All convolution layers specified in table 4.7 used kernels of size 3 by 3 and stride of 1. Besides that, all pooling layers used a kernel of size 2 by 2 and stride of 2.

4.2.3 ResNet50V2

Besides the *Simple* CNN, we also tried the *ResNet50V2* CNN, which is a more complex (deeper) architecture. The ResNet50V2 consist basically of several residual blocks (see section 2.2.1), composed with convolution and pooling layers (see section 2.2.2), stacked one after another. There is a lot of literature on the ResNet50V2 architecture ([He+15], [He+16]), so we will not dive into details here. However, the model architecture is

shown in table 4.8. The raw *ResNet50V2 feature extraction*, represent the feature extraction layers (i.e., all the layers containing convolution and/or pooling layers) of the ResNet50V2[3], while the remaining rows represent the layers intended to make the final prediction. The layers are evaluated from top to bottom. This model has $24,171,777$ free (learnable) parameters in total and it was trained with a learning rate of 0.0005.

Layer	Output Shape	Number of parameters
Input	$(bs, 224, 224, 38)$	0
Channel filtering	$(bs, 224, 224, 33)$	0
ResNet50V2 feature extraction	$(bs, 7, 7, 2048)$	23,612,672
Global Average Pooling	$(bs, 2048)$	0
Dense	$(bs, 256)$	524544
Batch Normalization	$(bs, 256)$	1024
ReLU	$(bs, 256)$	0
Dense	$(bs, 128)$	32896
Batch Normalization	$(bs, 128)$	512
ReLU	$(bs, 128)$	0
Dense	$(bs, 1)$	129

Table 4.8: ResNet50V2 CNN architecture. The rows represent each layer of the model. The flow of the model is from top to bottom. The *bs* on the *Output Shape* column stands for *Batch size*.

4.2.4 Xception

As we saw in section 2.2.1, each kernel in a regular convolution layer needs to simultaneously learn spatial and cross-channel correlations. For this reason, we also tested an architecture capable of separating these two tasks, the *Xception* [Cho17].

The Xception architecture combines the idea behind the Inception module and the residual blocks (see section 2.2.1). We will not dive into details about the Xception here. However, the model architecture is shown in table 4.9. The raw *Xception feature extraction*, represent the feature extraction layers (i.e., all the layers containing convolution and/or pooling layers) of the Xception[4], while the remaining rows represent

[3]For this work, we did not implement the ResNet50V2 architecture from scratch, instead we used the pre-built model that is provided in the Keras library. For more information pleases refer to the official documentation.

[4]For this work, we did not implement the Xception architecture from scratch, instead we used the pre-built model that is provided in the Keras library. For more information pleases refer to the official documentation.

the layers intended to make the final prediction. The layers are evaluated from top to bottom. This model has $21,373,929$ free (learnable) parameters in total and it was trained with a learning rate of 0.0005.

Layer	Output Shape	Number of parameters
Input	$(bs, 224, 224, 38)$	0
Channel filtering	$(bs, 224, 224, 33)$	0
Xception feature extraction	$(bs, 7, 7, 2048)$	20,814,824
Global Average Pooling	$(bs, 2048)$	0
Dense	$(bs, 256)$	524544
Batch Normalization	$(bs, 256)$	1024
ReLU	$(bs, 256)$	0
Dense	$(bs, 128)$	32896
Batch Normalization	$(bs, 128)$	512
ReLU	$(bs, 128)$	0
Dense	$(bs, 1)$	129

Table 4.9: Xception CNN architecture. The rows represent each layer of the model. The flow of the model is from top to bottom. The *bs* on the *Output Shape* column stands for *Batch size*.

4.2.5 Performance metrics

To evaluate and compare the performance of the models, besides the loss function (*Huber loss*), we also used 2 other error measures

- The Mean Squared Error (MSE)

$$E_{MSE}(Y, \hat{Y}) := \frac{1}{N}\sum_{n=1}^{N}(y_i - \hat{y}_i)^2 \tag{4.3}$$

- The Mean Absolute Error (MAE)

$$E_{MAE}(Y, \hat{Y}) := \frac{1}{N}\sum_{n=1}^{N}|y_i - \hat{y}_i| \tag{4.4}$$

where $Y, \hat{Y} \in \mathbb{R}^N$ are the true and predicted Transcription Rate (TR) values respectively.

Besides the error measures, we also used the *Coefficient of determination*[5] R^2, as well as the mean and standard deviation of the model error ($\bar{e}$ and $s(e)$ respectively), as a performance measures

[5]Intuitively, the *Coefficient of determination* R^2 represents how much of the variance in the target variable y is explained by the model when it is compared to $\bar{y}$ [SJ+60].

$$
\begin{aligned}
R^2 &:= 1 - \frac{SS_{res}}{SS_{tot}} \\
&:= 1 - \frac{\sum_{n=1}^{N}(y_i - \hat{y}_i)^2}{\sum_{n=1}^{N}(y_i - \bar{y})^2}
\end{aligned}
$$

$$
\begin{aligned}
\bar{e} &:= \frac{1}{N}\sum_{n=1}^{N}(y_i - \hat{y}_i) \\
&:= \frac{1}{N}\sum_{n=1}^{N} e_i
\end{aligned}
\tag{4.5}
$$

$$
s(e) := \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}(e_i - \bar{e})^2}
$$

where SS_{res} and SS_{tot} are the *Residual sum of squares* and the *Total sum of squares* respectively.

4.3 Interpretability methods

There are several hyper-parameters that need to be chosen in order to compute the score map for each cell image.

For the Integrated Gradient (IG) attribution map, recall that in practice computing ϕ^{IG} could be unfeasible or computationally very expensive. However, we can approximate ϕ^{IG} by means of $\phi^{Approx\ IG}$ (see equation 2.16). Therefore, we need to define the number of steps m for the Riemann sum approximation. In section 2.3.1 we also mentioned the necessity to set a baseline image x', which should contain no information about the image, in order to compute the IG. There are several options that can be used, each one of them with different advantages and disadvantages. However, for this work we only implemented two of them: 1) a simple black image (image containing only zeros) and 2) an image filled with Gaussian noise ($\mu = 0,\ \sigma = 1$). A very good analysis on the choice of the baseline can be found in this reference [SLL20].

In section 2.3.2 we saw that for VarGrad (VG) we need to define two parameters, the number of noisy images n (sample size) and the standard deviation σ for the the noise distribution.

As a rule of thumbs, a sample should not be smaller than 30, so this could be a feasible option. However, since Smilkov et al. [Smi+17] showed empirically that no further improvemnt (less noise) in score maps is observed for sample sizes greater than 50, we chose this bound as sample size.

Table 4.10 shows a summary of the parameters chosen to calculate the VarGrad Integrated Gradient (VGIG) score maps.

Method	Hyperparameter	Value
IG	m	70
	x'	black image
VG	n	50
	σ	1

Table 4.10: Parameters to compute score maps.

In section 2.3.1, we mentioned that the IG algorithm holds the *Completeness Axiom*, which means that the sum of all the components of the IG attribution map must be equal to the difference between the model's output evaluated at the image and the model's output evaluated at the baseline (see equation 2.15). This property allow us to check empirically if the number of steps m selected for the Riemann sum approximation is sufficiently large. Figure 4.3 shows that for the *simple CNN* model, a random image and $m = 70$, the completeness axiom is satisfied sufficiently well.

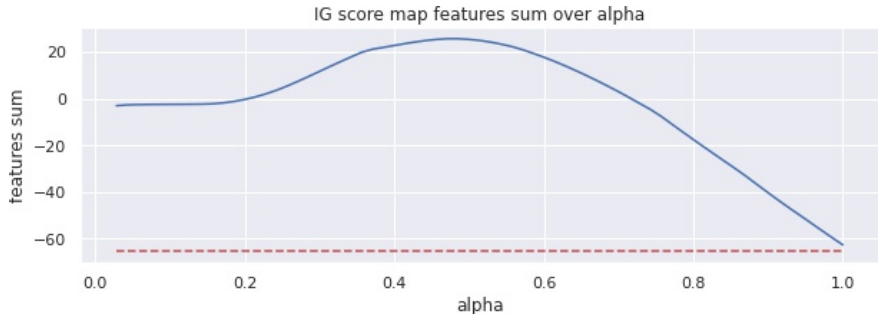


Figure 4.3: Sanity check for the number of steps m in the Riemann sum to approximate ϕ^{IG}. The red dotted line represent the difference $f(x) - f(x')$. The blue line represents the value of $\sum_i \phi_i^{Approx\ IG}(f, x, x', m)$ over α.

4.3.1 Discussion

For all the implemented architectures, L_1 and L_2 regularization was tried for both dense and convolution layers. However, in practice this did not significantly improve the generalization of the models, but it did increase the number of hyperparameters that need to be tuned (which means more models to train). For this reason the use of regularization was discarded for all the models shown on this work (L_1 and L_2 regularization strength was set to 0).

For the *ResNet50V2* and *Xception* architectures, the use of pre-trained weights and biases to initialize the model parameters (transfer learning) was also tried[6]. However, since this did not improve significantly the performance of the models, its use was discarded.

[6]This pre-trained parameters were obtained from the Keras library (which were fitted using the ImageNet dataset [Rus+15]).

Chapter 5

Results

This chapter is mainly divided into two parts

1. The model performance analysis
2. The model interpretation analysis

In the first part, we show the results of the models introduced in section 4.2, and compare the performance of each model against the others based on the metrics introduced in section 4.2.5. Besides, the performance of each model is compared to the reference values of the metrics, which validate that the models are capable of learning something meaningful from the data. This section ends by analyzing more in-depth the performance of the linear model against one of the Convolutional Neural Networks (CNNs) models, and shows that the latter is still capable of predicting fairly well the Transcription Rate (TR), even after significantly reducing the pixel intensity information of each channel and the correlation between them (by means of the data augmentation techniques introduced in section 3.3 and 4.1).

On the other hand, the Model interpretation section focuses on analyzing the results obtained using the interpretability methods introduced in section 2.3. The analysis begins with the division of the cells into three levels of transcription (low, medium and high), and ends by analyzing how the model changes the areas of interest in the input image as the TR changes. Furthermore, the analysis shows that as the TR increases, the model relies more on regions of the nucleus that are directly related to the genesis of mature messenger RNA (mRNA), which shows that interpretation methods can be used as tools to discover unknown biological relationships when applied to black-box models like CNNs.

The results of both sections also show that it is possible to predict (to some extent) the TR of a cell, based mainly on spatial information within the nucleus.

5.1 Model performance

In this section we show the results of each model in terms of Transcription Rate (TR) prediction. In addition, we provide reference values that validate that the models are

capable of learning something meaningful from the data. At the end of the section we compare the validation errors between the models and reference values.

The models were trained in the high performance computing cluster of the Helmholtz Zentrum München. The training jobs were allocated in the cluster by means of a job scheduling system (*Slurm*[1]), which a assigned the job to a node with a NVIDIA Tesla V100 GPU (either with 16 or 32 GB of dedicated ram memory.).

5.1.1 Baseline values for performance metrics

If we assume that there is no information in the training data $\boldsymbol{X}_{\text{train}}$ that can be used to explain the independent variable y (i.e. the TR), then a model f should not be able to make (learn) a better predictions than the average of the target values in the training set, i,e,

$$f(\boldsymbol{x}_i) = \bar{y}_{\text{train}} \tag{5.1}$$

for all $\boldsymbol{x}_i \in \boldsymbol{X}_{\text{train}}$ and where $\bar{y}_{\text{train}} = \frac{1}{N_{\text{train}}} \sum_{i=1}^{N_{\text{train}}} y_i$.

This idea is similar to the *Coefficient of determination* R^2, which represents how much of the variance in the target variable y is explained by the model when compared to $\bar{y}$. Thus, a model that always returns $\bar{y}$, will have a $R^2 = 0$. On the other hand, if the model gives better predictions than $\bar{y}$, then $0 < R^2 \leq 1$, otherwise $R^2 < 0$.

Table 5.1 shows the value of the performance measures (baseline values) (see section 4.2.5), when they are evaluated in the test set assuming that equation 5.1 holds.

$\bar{e}$	$s(e)$	R^2	MAE	MSE	Huber
4.86	59.99	-0.01	45.56	3622	45.07

Table 5.1: Baseline values for performance metrics evaluated in the test set.

As we already mentioned, the objective of this work is to explain cell expression using spatial information in multichannel images of cell nucleus. Therefore, it is important to estimate how much information related with the pixel intensities (color information) remains in the data after preprocessing and augmentation techniques. Since the linear model cannot take advantage of spatial information, we can use the performance of the linear model and the information in table 5.1 to estimate this. Then, if the linear model can reach a lower value in the loss function than the one shown in the table 5.1, this means that there is still color information in the data that can be used to predict the TR.

[1]Slurm is an open source, fault-tolerant, and highly scalable cluster management and job scheduling system for Linux clusters. For more information, please refer to the official documentation.

On the other hand, we can also use the information provided in table 5.1 to validate whether the Convolutional Neural Networks (CNNs) are capable of predicting the TR based only on spatial information.

5.1.2 Model performance comparison

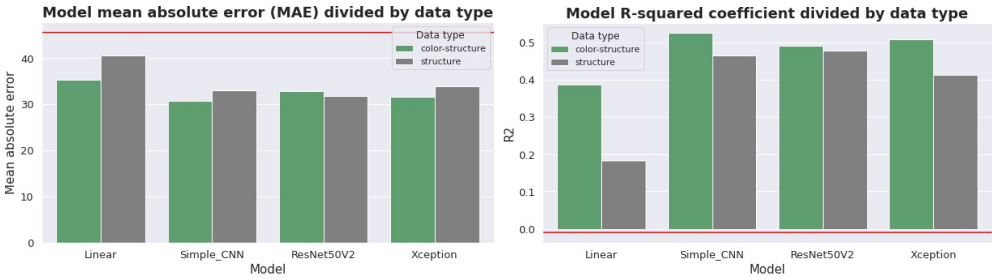
Table 5.2 shows a comparison of the performance of each model on the test set. Column *Data type* specifies the information contained in the training data; *structure* indicates only spatial information (which means that per-channel random color shifting was applied as augmentation technique to reduce pixel intensity information), while *color and structure* indicate spatial and pixel intensity information (which means that no random color shifting was applied). The row $\bar{y}$ (in the *Model* column) contains the baseline values for the performance metrics (see section 5.1.1). The numbers in bold indicate the models with the best overall performance (i.e., trained using data with and without pixel intensity information) per metric, while the shaded cells indicate the models with the best performance using only spatial data (structure).

Model	Data type	$\bar{e}$	$s(e)$	R^2	MAE	MSE	Huber
$\bar{y}$ (baseline)	targets avg	4.86	59.99	-0.01	45.56	3622	45.07
Linear	color-structure	4.06	46.83	0.38	35.26	2203	34.77
	structure	4.03	54.15	0.18	40.52	2941	40.02
Simple CNN	color-structure	3.00	**41.27**	**0.52**	**30.68**	**1708**	**30.18**
	structure	0.77	43.94	0.46	33.08	1926	32.59
ResNet50V2	color-structure	1.49	42.81	0.49	32.73	1830	32.24
	structure	**0.45**	43.38	0.47	31.83	1877	31.33
Xception	color-structure	6.69	41.57	0.50	31.66	1768	31.16
	structure	7.23	45.50	0.41	33.92	2117	33.42

Table 5.2: Model performance comparison. Performance measures where taken from the test set, with and without pixel intensity information (color-structure and structure respectively). Bold cells indicate the model-metric with best general performance. Shaded cells indicate the model-metric with best performance using only spatial (structure) data.

Figure 5.1 shows a graphical representation of the Mean Absolute Error (MAE) and R^2 values shown in table 5.2.

Table 5.2 shows that, as expected, for both types of training data (color-structure and structure) all the Convolutional Neural Network (CNN) models performed better than the linear model in all the performance measures, except for average error $\bar{e}$. Also, for all the error measures and the R^2 coefficient, both the linear model and the CNN models performed better than $\bar{y}$ (baseline values), which means that the models were able to learn something meaningful from both types of data. Surprisingly, the



(a) Model MAE comparison divided by data type.

(b) Model R^2 coefficient comparison divided by data type.

Figure 5.1: Graphic representation of the data shown in table 5.2, for the MAE and R^2 performance measures. Each group of bars represent a different model. The bar colors represent the data type used to train the models. The horizontal red line shows the baseline value.

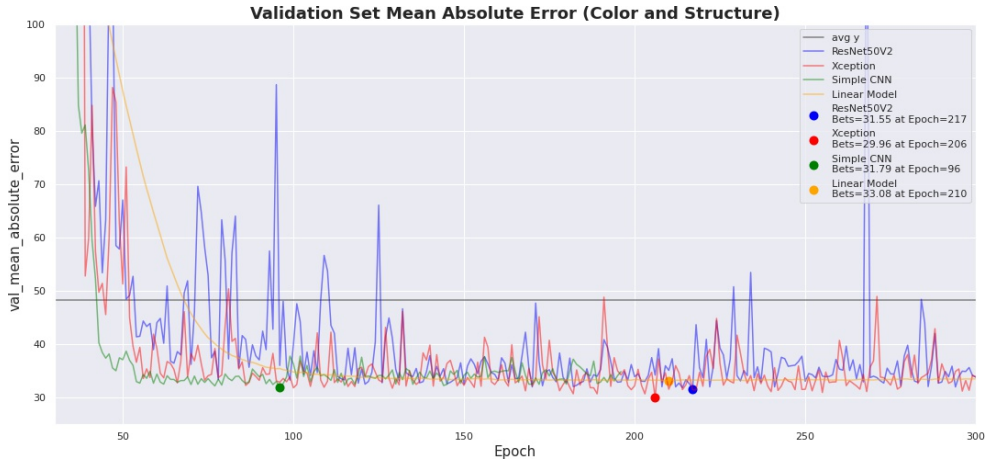
ResNet50V2 model was the only one that had a better performance in the structure data.

In general the *simple CNN* was the model with the best performance, while the *ResNet50V2* was the model with the best performance in the structure data. However, it is worth mentioning that the *simple CNN* model stayed behind the *ResNet50V2* model in the structure data, surpassing the *Xception* mode. Nevertheless, the performance of the *simple CNN* model was similar to that of the more complex models during training. This can be seen in figures 5.2a and 5.2b, which show the validation MAE of each model during training. In these figures we can see that the simple model has visibly less variance than the other two CNN models, especially in figure 5.2a.

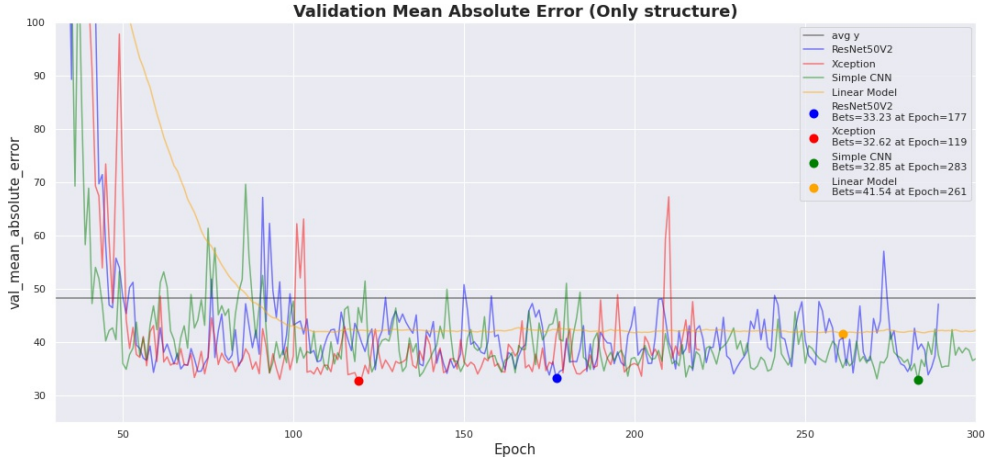
The dots in figure 5.2 indicate the epochs with the best performance with respect to the validation MAE of each model. The gray horizontal line corresponds to the MAE of the baseline evaluated in the validation set (see section 5.1.1). Due to early stopping, the number of epochs is not the same for all the models.

Figure 5.2b shows that the MAE of the linear model was generally higher than the MAEs of the CNN models, which reinforces our hypothesis that to some extent it is possible to describe cell expression, using only spatial information within the cell nucleus.

The *ResNet50V2* and *Xception* models have more than 24m and 21m of parameters respectively, while the *simple CNN* model has only around 160k. Therefore, the training of these two models require way more computational resources and time than the *simple CNN* model. However, table 5.2 and figure 5.2 show that the performance of the *simple CNN* model is similar to the *ResNet50V2* and *Xception*. Moreover, we observe that the importance maps of the *simple CNN* model (shown in section 5.2), were less noisy and informative than those obtained with the more complex models.



(a) Validation MAE using data with color and structure.



(b) Validation MAE using data only with structure.

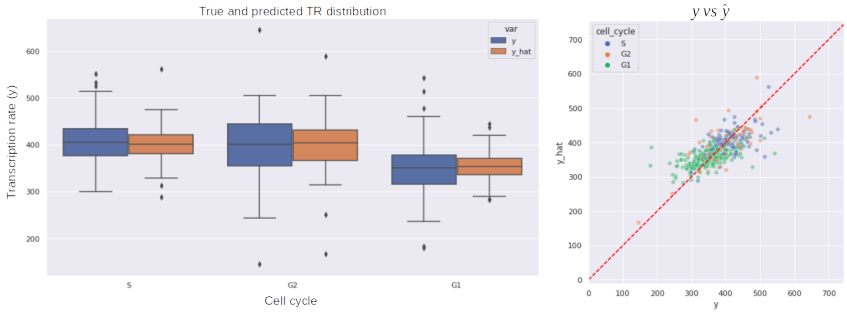
Figure 5.2: Validation MAE during training using data with (figure 5.2a) and without (figure 5.2b) pixel intensity information (color-structure and structure respectively). Each color represent a different model. The dot indicates the epoch in which the model reached its lowest validation MAE. The gray line indicates the baseline MAE in the validation set.

For this reason, in subsequent sections we will focus on the *simple CNN* model only.

5.1.3 Linear model

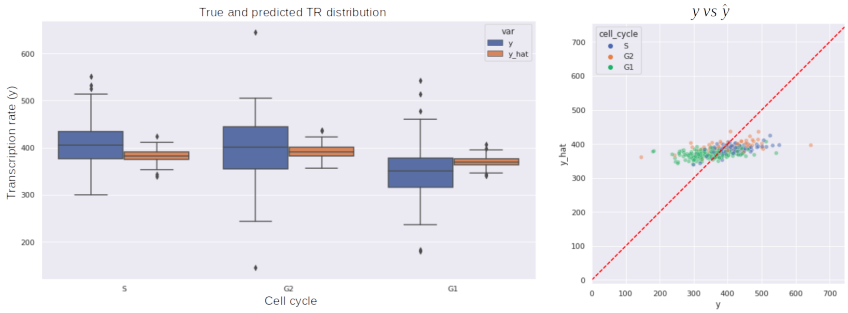
Figure 5.3 provides a more in-depth look at what was mentioned in section 5.1, with respect to the linear model. All the subfigures in 5.3 correspond to the test set and are

divided/colored by cell cycle.



(a) Distribution of y and $\hat{y}$ (boxplots) for data with color and structure.

(b) y vs. $\hat{y}$ for data with color and structure.



(c) Distribution of y and $\hat{y}$ (boxplots) for data with spatial information (structure only).

(d) y vs. $\hat{y}$ for data with spatial information.

Figure 5.3: Comparison between the true and predicted Transcription Rate (TR) (y and $\hat{y}$ respectively) for the linear model on the test set, divided by cell cycle. The first row of figures corresponds to the linear model trained with data containing pixel intensity and spatial information (color and structure), while the second row to the linear model trained with spatial data only (structure). The boxes in figures a and c show the first and third quartiles of the data (25% and 75% respectively), while the whiskers extend to show the rest of the distribution, except for points that are determined to be *outliers* using a function of the inter-quartile range. The line inside the boxes shows the second quartile (median) of the data. Figures b and d shows the true vs. predicted TR.

Subfigures 5.3b and 5.3d show that after removing the pixel intensity information from the training data (by applying per-channel random color shifting, see section 3.3), the linear model is unable to use the remaining spatial information, so it practically learns a constant function (similar to the average TR of the training set, see section 5.1.1). This can also be seen in subfigures 5.3a and 5.3c, which show a comparison between the true and predicted TR distributions. However, subfigures 5.3a and 5.3c

also show that even after reducing the pixel intensity information, the model was still able to learn slightly different average TRs for each cell cycle, which explains why the prediction of the linear model trained only with structure data is still slightly better than the baseline value (see table 5.2).

5.1.4 Simple CNN

Figure 5.4 provides a more in-depth look at what was mentioned in section 5.1, with respect to the *simple CNN* model. All the subfigures in 5.4 correspond to the test set and are divided/colored by cell cycle.

Unlike the linear model, subfigures 5.4b and 5.4d show that the simple CNN model training with structure data was able to approximate a function almost as good as that of the model trained with color and structure data. This can also be seen in subfigures 5.4a and 5.4c, which show that even after removing the pixel intensity information from the data, the model was still able to approximate fairly well the TR distribution.

Again, this reinforces our hypothesis that it is possible to describe cell expression to some extent, based solely on spatial information from the cell nucleus.

5.2 Model interpretation

In section 2.3 and 4.3 we explained the interpretability methods Integrated Gradient (IG) and VarGrad (VG) and how we can combine them to generate *score maps*, which can be used to identify model-important pixels in the input image.

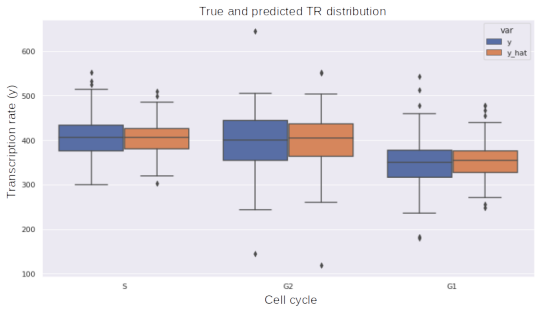
In this section we will analyze the results of the *simple CNN* model, which produced the most interesting score maps. However, we will not only investigate the areas of the input image that the model looks at to make a prediction, we will also analyze the dynamics of the score maps with respect to changes in the Transcription Rate (TR). In order to do this, we first introduce the methodology used to group cells by level of transcription.

This analysis results in the formulation of hypotheses about what the model focuses on to make its prediction as the transcription level increases, which can potentially indicate unknown biological factors that influence a cell's transcription.

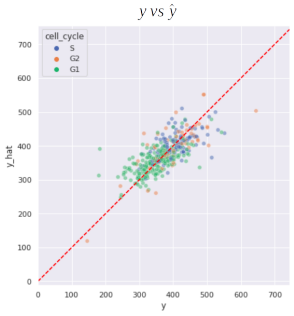
5.2.1 Cell grouping by transcription level

To divide the cells into different transcription level groups, we use one standard deviation away from the mean TR to classify a cell as low, medium or high transcription level. This is shown in table 5.3.

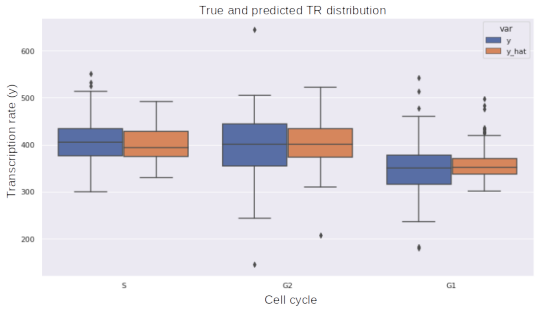
Table 5.3 also shows the number of cells and the percentage of data belonging to each group. Figure 5.5 shows the distribution of the TRs, as well as the division lines of each group (in red) and the average TR (in gray).



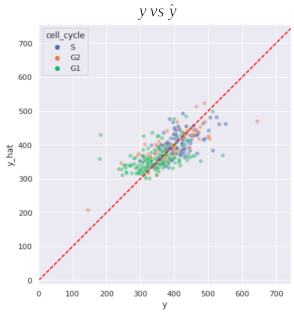
(a) Distribution of y and $\hat{y}$ (boxplots) for data with color and structure.



(b) y vs. $\hat{y}$ for data with color and structure.



(c) Distribution of y and $\hat{y}$ (boxplots) for data with spatial information (structure only).



(d) y vs. $\hat{y}$ for data with spatial information.

Figure 5.4: Comparison between the true and predicted Transcription Rate (TR) (y and $\hat{y}$ respectively) for the *simple CNN* model on the test set, divided by cell cycle. The first row of figures corresponds to the linear model trained with data containing pixel intensity and spatial information (color and structure), while the second row to the linear model trained with spatial data only (structure). The boxes in figures a and c show the first and third quartiles of the data (25% and 75% respectively), while the whiskers extend to show the rest of the distribution, except for points that are determined to be *outliers* using a function of the inter-quartile range. The line inside the boxes shows the second quartile (median) of the data. Figures b and d shows the true vs. predicted TR.

Figure 5.6 shows 3 cell nucleus images from the test set, randomly sampled from each of the 3 transcription groups. Each of the cells also corresponds to one of the 3 cell phases (G_1, S and G_2). The cells are shown from lowest to highest TR (from left to right). The images are the composition of channels *RB1_pS807_S811, PABPN1* and *PCNA* (for more information about this markers, please see tables A.3 and B.1).

Group	Grouping criterion	Criterion values	Group size	Group percentage
Low TR	$y \leq \bar{y} - s(y)$	$y \leq 316.56$	532	14.3%
Medium TR	$\bar{y} - s(y) < y < \bar{y} + s(y)$	$316.56 < y < 438$	2627	71%
High TR	$\bar{y} + s(y) \leq y$	$438 \leq y$	544	14.7%

Table 5.3: Cell grouping criteria with respect to their TR. The TR of each cell is denoted by y, while the mean TR and standard deviation by $\bar{y}$ and $s(y)$, respectively.

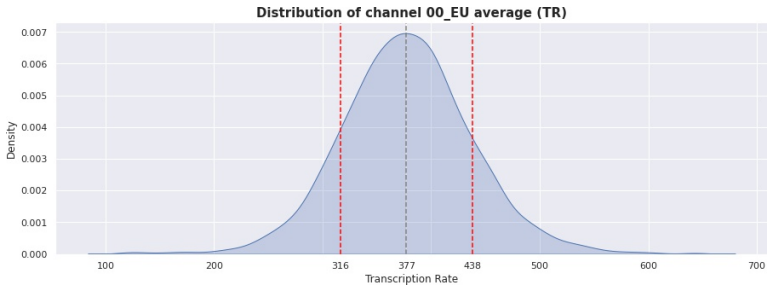
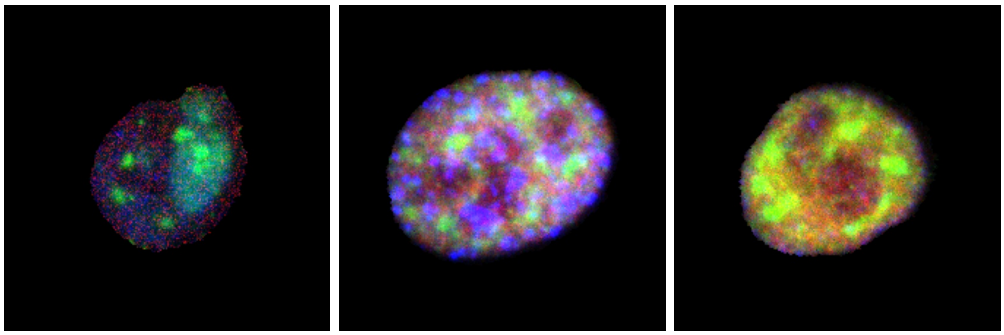


Figure 5.5: TR distribution. The red lines show the division between TR level groups. The gray line shows the mean TR.



(a) Sample cell 1 (cell id 277417); $y = 133.04$; TR group: *Low TR*; cell phase: G_1.

(b) Sample cell 2 (cell id 321001); $y = 378.19$; TR group: *Medium TR*; cell phase: S.

(c) Sample cell 3 (cell id 195536); $y = 540.09$; TR group: *High TR*; cell phase: G_2.

Figure 5.6: Cell nucleus images sample. The images are the composition of channels *RB1_pS807_S811*, *PABPN1* and *PCNA*. For each image, the TR is denoted by y.

5.2.2 Simple CNN model score maps

In this section we analyze the score maps corresponding to the *simple CNN* model trained using the data with spatial information (structure) only.

Figure 5.8 show the score maps corresponding to the cells in figure 5.6. This figure shows the channels *POL2RA_pS2, GTF2B, SRRM2, NCL, PABPN1, SETD1A, SON* and *SP100* (see tables A.3 and B.1) corresponding to the the score maps and the original cell nucleus images. Figure 5.8 also shows the overlap between the original image and its score map.

As we explained in section 2.3, a score map shows how important a pixel is to the output of the model with respect to the input image. Since the score map has the same shape as the input, we can see how important is a pixel at a per-channel level. We can generalize this idea of importance from a per-pixel-channel level to a per-channel level only, by summing all the pixel values corresponding to each channel.

To make the channel scores comparable between images, we normalize the scores by dividing them by the sum of all the pixels corresponding to the image score map. Therefore, the score of each channel will be a number between 0 and 1, and their sum will be always equal to 1. For the score maps shown in figure 5.8, the importance of each channel is indicated as a percentage at the top of the second row.

Figure 5.7 shows the average scores of the channels divided by level of transcription. The data correspond to the images belonging to the test set and the *simple CNN* model trained with spatial data only. The line at the top of each bar shows the 99% confidence interval for the mean channel importance.

In figure 5.7 we can see that the most important channels are the *POL2RA_ps2, GTF2B, SRRM2, NCL, PABPN1, COIL* and *SETD1A*, which accumulate more than 30% of the per-channel importance. Accordingly to table B.1 (in appendix B.1), the markers corresponding to these channels are aimed to

1. POL2RA_ps2: is an antibody that binds to the largest subunit of the RNA polymerase II (which is the enzyme responsible for transcribing DNA into pre-messenger RNA (pre-mRNA)) [Nov].

2. GTF2B: is an antibody that binds to the general transcription factor involved in the formation of the RNA polymerase II preinitiation complex [Lew04].

3. SRRM2: is an antibody that binds to a protein that in humans is required for pre-mRNA splicing as component of the *spliceosome*[2]. Along with the protein SON, SRRM2 is essential for Nuclear Speckles (NS)[3] formation [Ili+20].

4. NCL: is an antibody that binds to a protein that in humans is involved in the synthesis and maturation of ribosomes. It is located mainly in dense fibrillar regions of the nucleolus [ERA+88].

[2] A spliceosome is a large ribonucleoprotein complex found primarily within the nucleus of eukaryotic cells. The spliceosome removes introns from a transcribed pre-mRNA (see figure 2.4 on section 2.1.1) [WL11].

[3] The NS (also known as *Splicing speckles*) are structures inside the cell nucleus in which the pre-mRNA is transformed into a mature messenger RNA (mRNA) (see section 2.1.1) [SL11].

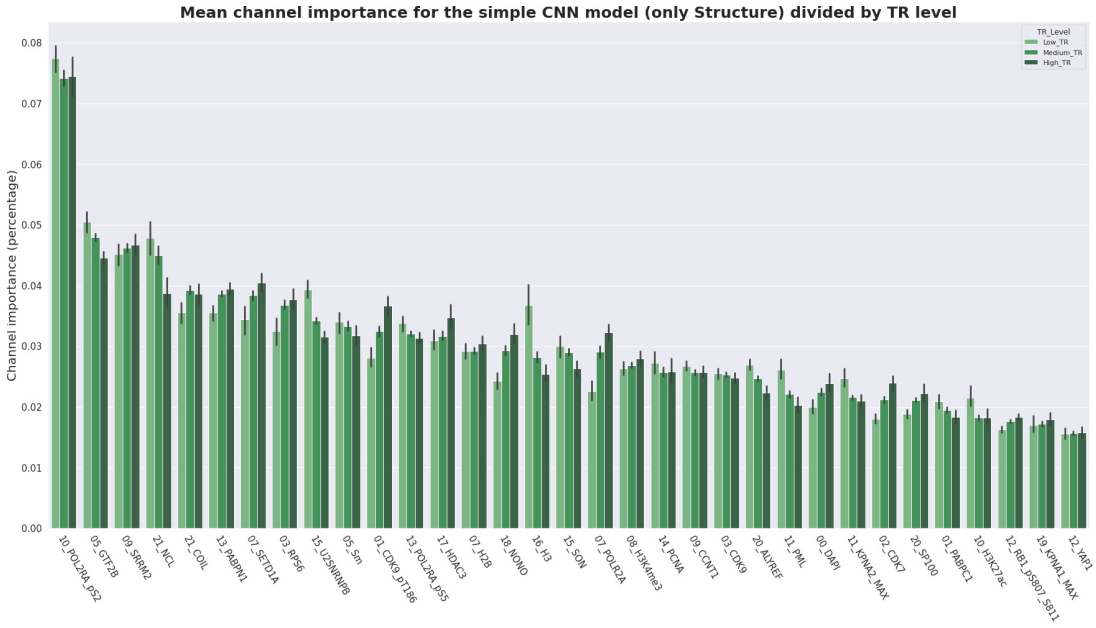


Figure 5.7: Average channel importance divided by transcription level corresponding to the *simple CNN model* trained with spatial data only. The data of the plot correspond to the images belonging to the test set. The 99% confidence interval for the mean channel importance is shown at the top of each bar.

5. PABPN1: is an antibody that binds to a protein involved in the addition of a Poly-A tail to the pre-mRNA during the splicing process (see figure 2.4 on section 2.1.1) [MDW15].

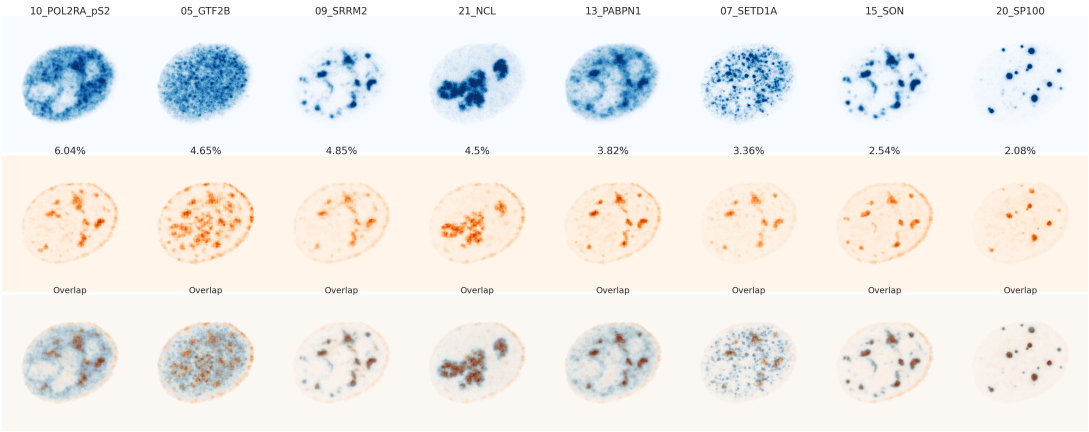
6. COIL: is an antibody that binds to a protein that is an integral component of *Cajal bodies*, which are nuclear suborganelles involved in the post-transcriptional modification of small nuclear and small nucleolar RNAs[Gena].

7. SETD1A: is an antibody that binds to a protein which is a component of a *histone*[4] methyltransferase (HMT) complex that produces mono-, di-, and trimethylated histone H3 at Lys4. Trimethylation of histone H3 at lysine 4 (H3K4me3) is a chromatin modification known to generally **marks the transcription start sites** of active genes [Genb].

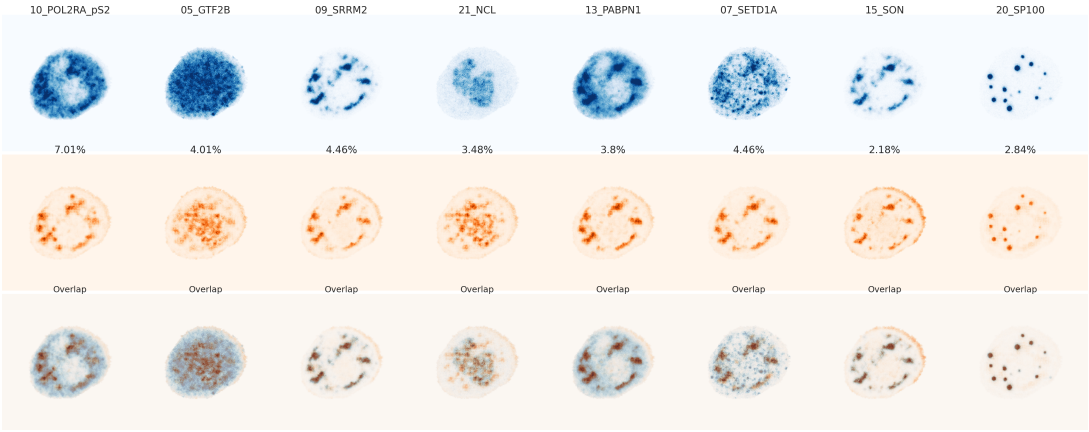
This means that these top channels are directly, or indirectly, related to the tran-

[4]A histone is a protein that provides structural support to a chromosome, so very long DNA molecules can fit into the cell nucleus. DNA molecules wrap around complexes of histone proteins, giving the chromosome a more compact shape [You06]. For a nice visualization of histone proteins, take a look at this link.

(a) Separate channels of the cell nucleus image and their score map corresponding to 5.6a.



(b) Separate channels of the cell nucleus image and their score map corresponding to 5.6b.



(c) Separate channels of the cell nucleus image and their score map corresponding to 5.6c.

Figure 5.8: Channels 25, 4, 7, 18, 11, 5, 13, 17 and 32 of score maps (corresponding to the *simple CNN model*) and cell images shown in figure 5.6. First row (in blue) shows the cell image, second row (in red) the score map (with the per-channel importance scores) and third row (blue and red) the overlap of the previous rows.

scription process. Particularly, the channels *POL2RA_ps2* and *GTF2B* are related to an early stage of the transcription process (i.e. the enzyme *RNA polymerase II*, which is essential to start the transcription process), while the channels *SRRM2* and *PABPN1* with the transcription process itself (the *splicing* process or maturation of pre-mRNA). The channel *SETD1A* is related to both pre-transcription and transcription processes, as this marks the transcription start sites of active genes. The channels *NCL* and *COIL* are not directly related with the transcription process. However, in figure 5.6 we can see that channel *NCL* indicates the nucleolus areas.

However, in figure 5.7 we can also see that as the TR grows, the top channels related to an early stage of the transcription process (*POL2RA_ps2* and *GTF2B*) lose relevance[5], while the top channels related to the splicing process (*SRRM2* and *PABPN1*) gain relevance. This suggests that the model relies on information with biological significance when predicting TR. Moreover, this suggests that as TR grows, the Convolutional Neural Network (CNN) goes from focusing on the synthesis of pre-mRNA to the synthesis of mature mRNA.

This statement is reinforced by looking at the score maps in figure 5.8. There we can see that as the TR increases (from figure 5.8a to 5.8c), the score maps corresponding to channels *POL2RA_ps2, SRRM2, PABPN1, SETD1A* and *SON* become more similar to the original *SRRM2* channel (which indicates the areas where the NS are).

5.2.3 Similarity between score maps and cell image

As we can see in figure 5.8, there are similarities between the score maps and the cell image channels. As we already mentioned, this suggest that the Convolutional Neural Network (CNN) is looking for specific information within the different image channels. Therefore, it is natural to ask ourselves which are the most *popular* (similar) cell image channels among the score maps.

To answer this question, for each cell image and its respective score map in the test set ($\boldsymbol{x}, \boldsymbol{s} \in X_{test} \subset \mathbb{R}^{D\times D\times C}$, respectively), we measure the similarity between the score map and the cell image channels. Then, for each score map channel we take its most similar cell image channel ($\boldsymbol{s}^i, \boldsymbol{x}^j \in \mathbb{R}^{D\times D}$, for $i, j \in \{1, \dots C\}$, respectively). Mathematically speaking, the cell image channel most similar to the score map channel $i \in \{1 \dots, C\}$ (denoted by $S_{min}(\boldsymbol{s}^i, \boldsymbol{x})$) is computed as follow

$$S_{min}(\boldsymbol{s}^i, \boldsymbol{x}) := \underset{j\in\{1\dots,C\}}{\arg\min}\{MAE(\boldsymbol{s}^i, \boldsymbol{x}^j)\} \tag{5.2}$$

where $MAE(\boldsymbol{s}^i, \boldsymbol{x}^j) := \frac{1}{D^2}\sum_{d_1=1}^{D}\sum_{d_2=1}^{D}|s^i_{d_1,d_2} - x^j_{d_1,d_2}|$ is the per-pixel mean absolute error between the score map channel i and a cell image channel j, with $i, j \in \{1 \dots, C\}$.

[5]The loss of relevance of channel *POL2RA_ps2* is not very clear in the plot 5.7 (corresponding to the test set). However, this trend is more noticeable in the plots corresponding to the training and validation sets.

Note that S_{min} always returns an index in $\{1 \ldots, C\}$.

Since we are only interested in measuring the similarity between score map channels and cell image channels at a spatial level (this means, not at a color or pixel intensity level), before applying 5.2 to the cell image and its respective score map, both are first standardized at a per-channel level[6].

Figure 5.9 shows the channels of the original images, divided by transcription level, which were selected as the most similar to the channels of the score maps.

The label above each bar represent the cumulative percentage of time (from left to right), the channels were selected as the most similar to one of the score maps channels. This cumulative percentage is also divided by transcription level.

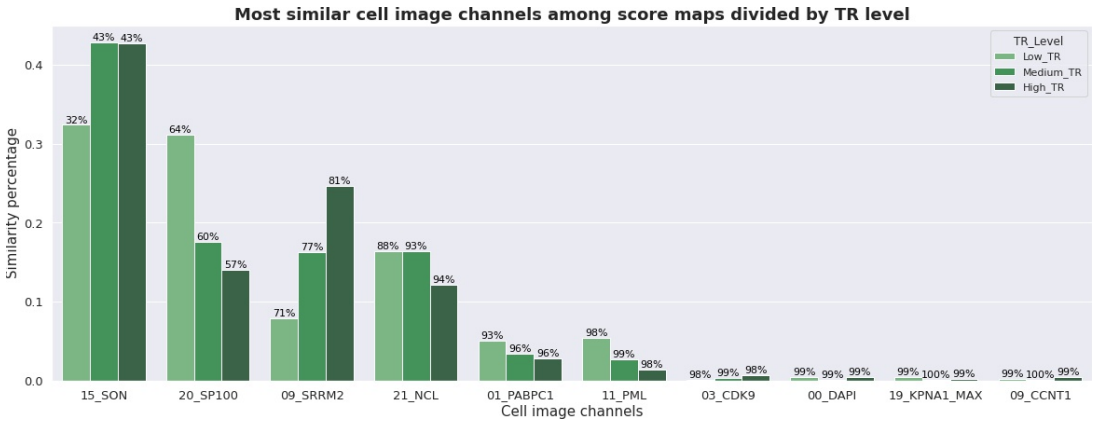


Figure 5.9: Top 10 most similar cell image channels to the score map channels divided by transcription level. The label above each bar represent the cumulative percentage of time that the channels were selected as the most similar.

In figure 5.9 we can see that around 98% of the times, only 6 channels (*SON, SP100, SRRM2, NCL, PABPC1* and *PML*) from the cell images are the most similar to the score maps channels, and half of this channels (*SRRM2, NCL* and *PABPC1*) are also in the top most active channels (see figure 5.7). Furthermore, figure 5.9 shows the same trend as that observed in figure 5.7, i.e. as the Transcription Rate (TR) increases, the channels related to the splicing process gain importance (*SON* and *SRRM2*), while the others lose it.

This is even more evident if we aggregate channels that target the same structures. This is, the *SON* and *SRRM2* channels, which indicate the areas where the Nuclear Speckles (NS) are; or the *SP100* and *PML* channels, which indicate the areas where the PML nuclear bodies are.

[6]Unlike as it was explained in section 3.2, where the standardization was done using parameters extracted from the training set, in this case the standardization parameters are calculated using the measured pixels of each channel (either from the cell image or its respective score map).

However, figure 5.9 only tell us the image channels that were the most similar to the score map channels in general. But, what if we would like to know this information at a per-channel level? Image 5.10 shows the cell images channels (columns) most similar to each score map channel (rows), divided by transcription level.

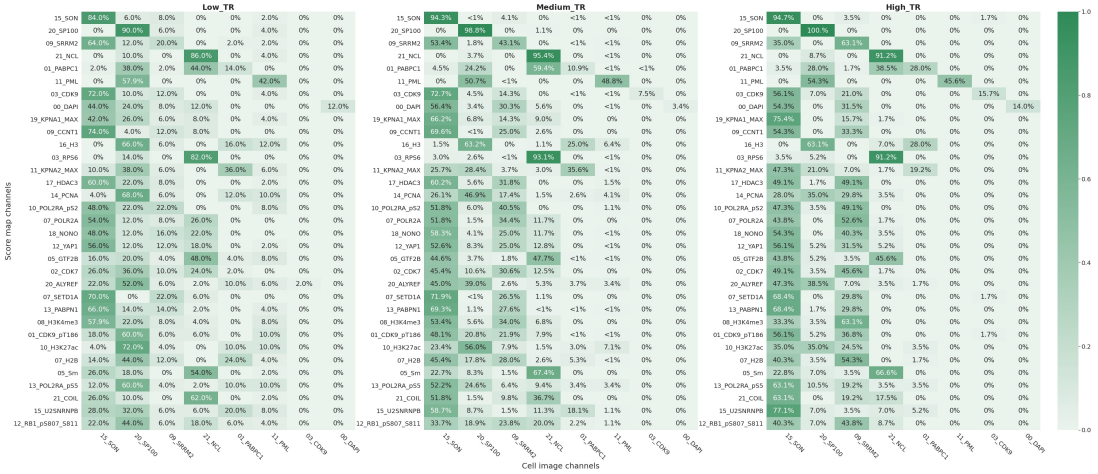


Figure 5.10: Most similar cell image channels to score map channels divided by transcription level.

As it was expected, figure 5.10 shows that most of the score map channels are similar to the same 6 cell image channels (*SON, SP100, SRRM2, NCL, PABPC1* and *PML*). However, this figure also shows that there are notable differences between the score map channels. Furthermore, it can be observed that as the TR increases, the score map channels become more similar to the *SON* and *SRRM2* channels of the cell images, which indicates the areas where NS are. For example, the score map channel *POL2RA_ps2*, which is the most active during the TR prediction (see figure 5.7), goes from being similar to the image channel *SP100* 22% of the time when the TR is low, to only 3.5% when the TR is high.

But what about the cell image channel *SP100*, which is very popular among score maps channels when the TR is low, but loses relevance when the TR is high? In figure 5.10 it is observed that the score map channel *SP100* is similar to its equivalent in the cell image 90% of the time when the TR is low. However, as the TR increases, this percentage grows up to 100%. This agrees with what is observed in image 5.7, which shows that this channel gains relevance in the prediction of TR as it increases.

Similarly, in figure 5.8 it can be seen that as the TR increases, the *SP100* channel shows more defined subnuclear organelles. Furthermore, it can be observed that these organelles are formed in the regions where the NSs are (indicated by channels *SRRM2* and *SON*). This is very interesting as it suggests that for the model the *SP100* channel (which is closely related to the *PML* channel, see table B.1) has a high influence on

the TR prediction. It also may suggests that cells with a high number of PML bodies in the cell nucleus, could be an indicative of a high transcription. However, this is a hypothesis that would have to be rigorously studied and validated.

Again, figures 5.9 and 5.10 reinforce our hypothesis that as the TR increases, the CNN goes from focusing on the synthesis of pre-messenger RNA (pre-mRNA) to the synthesis of mature messenger RNA (mRNA). Furthermore, figure 5.10 indicates that as the TR grows, the score map channels tend to be more similar to the *SRRM2* and *SON* cell image channel, which means that the CNN is simply looking for splicing signals in all the input image channels.

The results show that interpretability techniques can help us to understand how black-box models, like CNNs, make their predictions. This allows us to learn from models and has the potential to help us make biological discoveries.

5.3 Discussion

The models shown in this work were trained using the computing resources of the Institute of Computational Biology (ICB) of the Helmholtz Zentrum München (HMGU). Therefore, the training time of each model depends on the concurrency of the cluster and the node assigned by the job scheduling system (SLURM). For this reason, it is not possible to provide a fair comparison of the training time between models.

Since we are dealing with multichannel images, we may have highly correlated features in the data. For this reason, the model could focus on one or other features during training. For this reason, as a future work it would be interesting to train the models with different initializations, and validate that the score maps obtained are consistent. Another option could be to train the models aggregating correlated channels, based on the information obtained in the interpretability analysis.

Another validation method that was contemplated in this work, but that was not possible to complete, is the RemOve And Retrain (ROAR) [Hoo+18]. This method would have allowed us to observe to what extent the performance of the model would be degraded, by replacing the pixels marked as important by the interpretability methods by uninformative pixels. When comparing the increase in error against a random selection of pixels, this would show if the pixels selected as important were really better than just a random guessing. This could also help us to detect highly correlated features in the data.

Chapter 6

Conclusion and future work

6.1 Conclusion

In this work we use a dataset generated by means of the multiplexed protein map (MPM) protocol to predict the Transcription Rate (TR) of a cell, using spatial information from its nucleus and a Convolutional Neural Network (CNN). In order to do this, the first step was to reconstruct the multichannel images of cell nucleus from pixel intensity readouts stored in text files, and then build a dataset that allows to train models easily and efficiently. To reduce as much as possible the information contained in the intensity of the pixels and the correlation of these between the channels, different preprocessing and data augmentation techniques were tried. This allowed the CNN to focus mainly on the information contained in the location, distribution, shape and size of elements within the cell nucleus to predict its TR.

For this work different architectures were tried. From a *simple CNN* (with only 160k parameters), to more complex architectures such as the *ResNet50V2* and the *Xception* (with more than 20m parameters). Besides, for all the implemented architectures, L_1 and L_2 regularization was tried for both dense and convolution layers. However, in practice this did not significantly improve the generalization of the models, but rather increased the number of hyperparameters. For this reason the use of regularization was discarded from the final results. For the *ResNet50V2* and *Xception* architectures, the use of pre-trained weights and biases to initialize the model parameters (transfer learning) was also tried. However, since this did not significantly improve the performance of the models, its use was also discarded.

Overall, the results showed that it is possible to predict TR using only localization information from the cell nucleus, which is an interesting biological finding by itself. This means that the distribution of proteins in the nucleus is very important, and not only their overall abundance. This shows the importance of generating data with subnuclear localization information, and the developing of methods to evaluate this data.

However, predicting TR was not the only objective of this work. Through the interpretability methods Integrated Gradient (IG) and VarGrad (VG), for each cell nucleus image we obtained score maps that allowed us to observe the pixels that the

CNN considered as relevant to predict the TR. This allowed us to know which proteins and nuclear bodies were most relevant for the prediction of TR, and how this changes as the TR decreases or increases. This led us to formulate the hypothesis that as the TR grows, the CNN goes from focusing on the synthesis of pre-messenger RNA (pre-mRNA) to the synthesis of mature messenger RNA (mRNA).

In addition to this, the analysis of the dynamics of the score maps showed that for the model the *SP100* channel (which is closely related to the *PML* channel) has a high influence on the prediction of the TR. This may suggest that cells with a high number of PML bodies in the cell nucleus, could be an indicative of a high transcription. However, this is a hypothesis that would have to be rigorously studied and validated.

This shows that interpretability methods can help us to understand how CNNs makes its predictions. This is very important, as understanding how a model works has the potential to help us make scientific discoveries.

In order to reduce the noise in the score maps, the use of random noise in the input images during model training was also tried. However, this did not show any apparent improvement in the scores more, so its use was discarded.

For each of the three proposed architectures, the score map of each cell nucleus image was computed. However, the score maps corresponding to the *simple CNN* architecture were less noisy and showed more defined nuclear structures. For this reason, and besides the fact that the difference in performance against the more complex architectures was minimal, the *simple CNN* was selected to generate the results shown in section 5.2.

6.2 Future work

As we mentioned in chapter 3, for this work only the images corresponding to cells without metabolic or pharmacological perturbations were used. Therefore, it would be interesting to include all the available cells in the dataset and see how the different perturbations change the proteins and/or organelles indicated as important by the score maps.

In chapter 3 we also mentioned that the use of an unbalanced dataset (with respect to the number of cells in the G_1, S or G_2 phases) during training, could result in a bias model. For this reason, as a future work, a model could be trained using a balanced dataset and compare its performance with current results. Also, it would be interesting to see if this results in different score maps.

Because it has been observed that the *EU* marker could binds to DNA molecules if the incubation time is too long ([JS08] and [Bao+18]), as a future work another MPM dataset could be analyzed, either with a shorter or longer incubation time for the *EU* marker. Then, it would be interesting to validate if the results (in performance and score maps) obtained with both datasets (and the current one) are consistent.

We have seen that interpretability methods can help us understand how CNNs makes

its predictions, which has the potential to leverage scientific discoveries. For this reason, it would be interesting to explore other interpretability methods or investigate more IG, for example by using different baselines and see the impact this has on the score maps.

As we saw in chapter 5, the TR can be predicted using only localization information of proteins within the cell nucleus, and not just their overall abundance. Therefore, it would be interesting to study to what extent the information on the amount of proteins and their spatial information complement each other in the prediction of the TR, or to what extent they are correlated.

As we explained in section 2.3, the interpretability methods aim to rank the input pixels based on how much they contribute to the output of the model. However, there is no way to prove this mathematically. For this reason, as future work it would be necessary to implement a method that allows validating the veracity of the score maps obtained at least empirically. One possible option is the RemOve And Retrain (ROAR) methodology [Hoo+18], which would allow us to observe to what extent the performance of the model would be degraded, by replacing the pixels marked as important in the score maps with non-informative pixels. When comparing the increase in error with a random selection of pixels, this would show whether the pixels indicated as important by the interpretability methods were actually better than a simple random guess. This could also help us detect highly correlated features in the data.

Appendix A

Remarks on implementation

This appendix contains notes about how to execute all the scripts and notebooks used in this work. It also contains information about the parameters that need to be specified for each program. All the scripts and notebooks were written in `Python` and executed over `Anaconda`. You can find information about the environment setup, packages version, etc. here.

The logic to execute any `Python` script is always the same

```
python python_script_name.py -p ./Parameters_file_name.json
```

For the `Jupyter Notebooks` you just have to open it and set the variable `PARAMETERS_FILE` with the absolute path and name of the input parameters file

```
PARAMETERS_FILE = "/path_to_file_dir/Parameters_file_name.json"
```

For each script/notebook all the needed parameters have to be specified inside its parameter file only. The format for the parameters file is always `JSON` and the parameters values are specified in a python-dictionary format. Al Binary parameters need to be specified as 1 or 0 (True or False respectively).

A.1 Raw data processing and QC implementation notes

This appendix contains implementation and technical notes about the data preprocessing process that needs to be performed before the construction of the dataset used to train the models. This process is performed by a single Pyhton script (Jupyter Notebook) and contemplate two main steps

1. The reconstruction of single cell images from the raw data (text files).

2. The discrimination of single cell images accordingly to a quality control.

As it is explained in section 3.2, the protein readout of each well are contained in several files. Here we introduce those that are relevant for this work

- `mpp.npy`: 2D NumPy array. Each row contains the protein readouts (intensities) of each pixel of the well (one column per protein). The values of this array vary from 0 to 65535 i.e. 2^{16} i.e. 2 bytes or 16 bits.
- `x.npy`/`y.npy`: 1D NumPy array. Each entrance contains the x/y coordinate of a pixel of the well protein readouts (i.e. `x.npy` and `y.npy` map the protein readouts in `mpp.npy` with a 2D plane). Accordingly with [GHP18], the size of a single channel well image is 2560x2160. Therefore, the values in `x.npy` vary between 1 and 2560 and form 1 to 2160 for `y.npy`
- `mapobject_ids.npy`: 1D NumPy array. Each entrance contains an id that maps the protein readouts in `mpp.npy` with the nucleus of a cell in the well. Each cell nucleus in the well is identified by a unique id.

Since files `mpp.npy`, `x.npy`, `y.npy` and `mapobject_ids.npy` contains different parts of the well protein readouts, the first dimension of the arrays contained in the files always has the same size.

Beside the files with protein readouts (`npy` files), each well also comes with two additional `csv` files[1] containing further information about each cell in the well

- `metadata.csv`. Contains one raw per single cell nucleus in the well. The mapping between the metadata file and the protein readouts (`npy` files), is made through the column *mapobject_id*, which uniquely identify cells (but only within the well). On the other hand, column *mapobject_id_cell* uniquely identify cells across all wells. Columns *is_polynuclei_HeLa* and *is_polynuclei_184A1* indicate if a cell is in mitosis phase. This metadata file also contains information about the experimental setup, like plate name, well name, site position, etc..
- `channels.csv`. Contains only two columns that maps the immunofluorescence marker name (column *channel_name*) and the channel id (column *id*) of the protein readouts.

The files introduced so far are specific to each well. However, we still need to introduce other 3 files which contains information about all the wells

- `secondary_only_relative_normalisation.csv`. Contains the experimental setup information related to the image capturing process. Among other information, it contains the *background value* of each channel that has to be subtracted from the protein readouts during the reconstruction of the images.
- `cell_cycle_classification.csv`. Contains the phase of each cell.

[1] These `csv` files can be easily opened as a *Pandas DataFrame*. For more information, please refer to the official documentation.

- `wells_metadata.csv`. Contains more information about the experimental setup. Among other information, it contains the pharmacological/metabolic perturbation applied to each well.

To execute the raw data processing, one has to open the Python Jupyter Notebook `MPPData_into_images_no_split.ipynb` and replace the variable `PARAMETERS_FILE` with the absolute path and name of the input parameters file before running the notebook. A sample parameter file (`MppData_to_imgs_no_split_sample.json`) is provided along with this work. It contains the parameters used for the experiments shown in this work. Table A.1 provides an explanation of some of this parameters.

JSON variable name	Description
`raw_data_dir`	Path where the directories that contain the raw data files of each well are
`perturbations_and_wells`	Dictionary. The dictionary keys must be the directories for each perturbation, while the elements (a list) must contain the directory name of each well (one list entrance per well)
`output_pp_data_path`	Path where the output folder of the notebook must be located
`output_pp_data_dir_name`	Folder name where the notebook output will be saved
`img_saving_mode`	Indicate the shape of the output images. To replicate the experiments of this work, this variable must be set to `original_img_and_fixed_size`, which means squared images of fixed size
`img_size`	Integer. High and width of the output image (squared)
`return_cell_size_ratio`	Binary. Indicate if cell size ratio (percentage of the image that is occupied by the cell nucleus measurements) must be added to the output metadata file. During the data augmentation, this information can be used to approximate the parameters of the distribution used to randomly vary the size of the cell nucleus
`background_value`	Path and name (normally `secondary_only_relative_normalisation.csv`) of the metadata file containing the per-channel background values
`subtract_background`	Binary. Indicate if background color need to be subtracted from each channel

`cell_cycle_file`	Path and name (normally `cell_cycle_classification.csv`) of the metadata file containing the phase of each cell
`add_cell_cycle_to_metadata`	Binary. Indicate if cell phase must be add to the output metadata file
`well_info_file`	Path and name (normally `wells_metadata.csv`) of the metadata file containing the information about well perturbation
`add_well_info_to_metadata`	Binary. Indicate if columns of `well_info_file` must be add to the output metadata file
`filter_criteria`	List containing the metadata columns names that will be used in the quality control. For this work ["is_border_cell", "is_polynuclei_184A1", "is_polynuclei_HeLa", "cell_cycle"] was used
`filter_values`	List containing the filtered values for the columns indicated in `filter_criteria`. For this work [1, 1, 1, "NaN"] was used
`aggregate_output`	Indicate how to project each image channel into a number. Must be equal to "avg" (average)
`project_into_scalar`	Binary. Indicate if the channel scalar projection must be add to the output metadata file

Table A.1: Parameters to perform the raw data processing.

Roughly speaking, the notebook iterates over the specified wells sequentially. This means that for each well the notebook

1. Reads the well metadata file `metadata.csv` and merge it with the general metadata files, `cell_cycle_classification.csv` and `wells_metadata.csv`.
2. Performs the quality control and select the ids (*mapobject_id_cell*) that were approved.
3. Converts[2] and saves the selected ids using the well protein readouts files `mpp.npy`, `x.npy`, `y.npy`, `mapobject_ids.npy` and the general file `secondary_only_relative_normalisation.csv`.

The notebook also saves at the end a general metadata file (`csv` file), containing the metadata of all the processed wells.

[2]The library `mpp_data_V2.py` used to perform the raw data transformation, is almost entirely based on Dr. Hannah Spitzer library `mpp_data.py`. I thank the Dr. Spitzer for providing me with her library for this work.

A.2 TensorFlow dataset and image preprocessing implementation notes

After the raw data was processed and converted into images of single cell nucleus (see section 3 and appendix A.1), it is possible to build a TensorFlow Dataset (TFDS) data can easily end efficiently feed data into a model built in TensorFlow. A TensorFlow Dataset is build by writing a subclass of the `tensorflow_datasets.core.DatasetBuilder` class (for more information, please refer to the official documentation), and there are some steps that need to be followed to do so. The easiest way to build a TFDS, is by running the bash script `Create_tf_dataset.sh`, which executes this steps. The script needs to be executed (and located) in the same directory where the folder containing the Python code to build the dataset is

```
./Create_tf_dataset.sh -o /Path_to_save_TFDS -n ↷
    ↪ Folder_name_containing_the_TFDS_builder_code -p ↷
    ↪ /Path_to_parameters_files/parameters_file.json -e ↷
    ↪ my_conda_env_name
```

where the flag `-o` indicates the path where TFDS will be located after it is built, `-n` the name of the directory (not the path, the folder name in the same directory as the script) containing the python (builder) code for required dataset, `-p` the absolute path and name of the input parameters file[3] and `-e` the name of the Anaconda environment used to build the TFDS. The specified Anaconda environment is necessary not just to build the TFDS, but also to register it in the environment. If a TFDS is not registers in an Anaconda environment, the `tensorflow_datasets`[4] library will not find it, and the user will not be able to call it and use it. Therefore, to register a custom TFDS in another environment, one just have to execute the `Create_tf_dataset.sh` script specifying the new environment using the `-e` flag. If the TFDS was already built by another environment, python will just register the dataset under the new environment and it will not build it again.

Table A.2 provides an explanation of the variables contained in the parameters file.

JSON variable name	Description
`data_source_parameters`	Path where the parameters file used in the raw data processing is (see appendix A.1). Several parameters from this file are used to build the TFDS
`perturbations`	A list containing the names of the perturbations to be included in the TFDS. For instance, ["normal", "DMSO"]

[3]This file needs to be `JSON` format and located in a directory named `Parameters`, which needs to be located inside the directory specified by the `-n` flag.

[4]See the documentation here.

`cell_cycles`	A list containing the names of the cell phases to be included in the TFDS. For instance, ["G1", "S", "G2"]
`train_frac`	Scalar between 0 and 1. Proportion of the data to include in the train set
`val_frac`	Scalar between 0 and 1. Proportion of the data to include in the validation set. Proportion of the data to include in the test set is 1 - `train_frac` - `val_frac`
`seed`	Scalar. For reproducibility of the train, val and test split
`percentile`	Scalar between 0 and 100. Percentile used in clipping and/or linear scaling and/or standardization
`apply_clipping`	Binary. If 1, per-channel clipping is applied using the channel percentile
`apply_mean_extraction`	Binary. If 1, per-channel mean shift is applied using the channel mean
`apply_linear_scaling`	Binary. If 1, per-channel scaling is applied using the channel percentile
`apply_z_score`	Binary. If 1, per-channel standardization is applied using the channel parameters
`input_channels`	List containing the name of the channels (elements of the column *Marker identifier* of table A.3) to be included in the images contained in the TFDS
`output_channels`	List of only ONE element containing the name of the channel to be used as the target variable (its protection, i.e. the channel average)

Table A.2: Parameters to perform the raw data processing.

As it is shown in table A.2, the parameter `input_channels` specifies the channels that will be included in the TFDS images (see table A.3). However, to avoid building a new dataset every time we change the input channels, all the channels are included here and then filtered in the model (see section 3.2.3 for a more detailed explanation).

A sample parameter file (`tf_dataset_parameters_sample.json`) is provided along with this work. It contains the parameters used in the Python script `MPP_DS_normal_DMSO_z_score.py`, to build the dataset `mpp_ds_normal_dmso_z_score` used to train the models in this work.

Channel name	Marker identifier	Raw data id	TFDS id
DAPI	00_DAPI	0	0
H2B	07_H2B	1	1
CDK9_pT186	01_CDK9_pT186	2	2
CDK9	03_CDK9	3	3
GTF2B	05_GTF2B	4	4
SETD1A	07_SETD1A	5	5
H3K4me3	08_H3K4me3	6	6
SRRM2	09_SRRM2	7	7
H3K27ac	10_H3K27ac	8	8
KPNA2_MAX	11_KPNA2_MAX	9	9
RB1_pS807_S811	12_RB1_pS807_S811	10	10
PABPN1	13_PABPN1	11	11
PCNA	14_PCNA	12	12
SON	15_SON	13	13
H3	16_H3	14	14
HDAC3	17_HDAC3	15	15
KPNA1_MAX	19_KPNA1_MAX	16	16
SP100	20_SP100	17	17
NCL	21_NCL	18	18
PABPC1	01_PABPC1	19	19
CDK7	02_CDK7	20	20
RPS6	03_RPS6	21	21
Sm	05_Sm	22	22
POLR2A	07_POLR2A	23	23
CCNT1	09_CCNT1	24	24
POL2RA_pS2	10_POL2RA_pS2	25	25
PML	11_PML	26	26
YAP1	12_YAP1	27	27
POL2RA_pS5	13_POL2RA_pS5	28	28
U2SNRNPB	15_U2SNRNPB	29	29
NONO	18_NONO	30	30
ALYREF	20_ALYREF	31	31
COIL	21_COIL	32	32
BG488	00_BG488	33	33
BG568	00_BG568	34	34
EU	00_EU	35	NA
SRRM2_ILASTIK	09_SRRM2_ILASTIK	36	35
SON_ILASTIK	15_SON_ILASTIK	37	36
Cell mask	NA	NA	37

Table A.3: Image channels. Column *Raw data id* shows the channel id used in the raw data, while column *TFDS id* shows the channel id used in the TensorFlow dataset.

A.3 Model training implementation notes

This appendix is intended to provide a brief explanation of how to run the Python script (Jupyter Notebook) responsible for training the models used in this work. In addition, here we also provide a short explanation of the parameter file that must be specified to train any model.

Since data augmentation techniques can be selected independently for each trained model, their corresponding hyperparameters are also explained here.

The Jupyter Notebook responsible for training the models is the one that requires the largest number of parameters. However, the function `set_model_default_parameters` (in the `Utils.py` library) provides default values for all the parameters. Therefore, if some hyperparameter is not specified here or in section 4.2, then the value used was the one specified in that function.

To train a model, one has to open the Python Jupyter Notebook `Model_training_class.ipynb` and replace the variable `PARAMETERS_FILE` with the absolute path and name of the input parameters file before running the notebook. A sample parameter file (`Train_model_sample.json`) is provided along with this work. It contains the parameters used to train the *Simple* Convolutional Neural Network (CNN) (see section 4.2.2), using the data augmentation techniques specified in section 4.1.2. Table A.4 provides an explanation of some of the model training parameters, while table A.5 an explanation of some of the data augmentation parameters. Although the training and data augmentation parameters are specified in separate tables, they must be in the same `JSON` parameter file (and also as items in the same dictionary).

JSON variable name	Description
`model_name`	Name of the architecture to be trained. Available: `simple_CNN`, `ResNet50V2`, `Xception`, `Linear_Regression`
`pre_training`	Binary, whether or not use pretrained weights and biases as initial parameters. Only available for `ResNet50V2` or `Xception` architectures
`dense_reg`	$[L_1, L_2]$, where L_1 and L_2 are the regularization strengths for the dense layers weights
`conv_reg`	$[L_1, L_2]$, where L_1 and L_2 are the regularization strengths for the convolution layers weights
`bias_l2_reg`	L_2 regularization strengths for convolution and dense layers biases
`number_of_epochs`	Maximum number of epochs to train
`early_stop_patience`	For early stopping. Specify how many epochs at most the model can train without decreasing the loss function before stopping the training

`loss`	Loss function name. Available: `mse`, `huber`, `mean_absolute_error`
`learning_rate`	Learning rate for Adam optimizer
`BATCH_SIZE`	Batch size
`model_path`	Path to save the models and checkpoints
`clean_model_dir`	Binary, whether or not to delete the content of the directory specified by `model_path`
`tf_ds_name`	Name of the TFDS to be used during training
`local_tf_datasets`	Local path where the TFDSs are stored
`input_channels`	List containing the name of the channels (elements of the column *Marker identifier* of table A.3) to be included in the images contained in the TensorFlow Dataset (TFDS)
`shuffle_files`	Binary, whether or not to shuffle the dataset at the beginning of each epoch
`seed`	Random seed to reproduce the shuffling of the TFDS

Table A.4: Model training parameters.

JSON variable name	Description
`random_horizontal_flipping`	Binary, whether or not to perform random horizontal flips on the training set
`random_90deg_rotations`	Binary, whether or not to perform random 90deg rotations on the training set
`CenterZoom`	Binary, whether or not to perform random center zoom-in/out on the training set
`CenterZoom_mode`	Zoom proportion R.V. distribution. Available: `random_normal`, `random_uniform`
`Random_channel_intencity`	Binary, whether or not to perform per-channel random color shifting on the training set
`RCI_dist`	Distribution of random color shifts. Available: `uniform`, `normal`. If `uniform` distribution selected ($U(-a, a)$), then $a = \mu + 3\sigma$
`RCI_mean`	Mean μ for the distribution specified by `RCI_dist`
`RCI_stddev`	Standard deviation σ for the distribution specified by `RCI_dist`
`Random_noise`	Binary, whether or not to add random normal noise ($N(0, \sigma)$) on the training set
`Random_noise_stddev`	Standard deviation corresponding to the normal distribution of random noise

Table A.5: Data augmentation parameters.

A.4 VarGrad IG implementation notes

In order to generate the VarGrad Integrated Gradient score maps, you must execute the python script `get_VarGradIG_from_TFDS.py` specifying the parameters file

```
python get_VarGradIG_from_TFDS.py -p ./Parameters_file_name.json
```

Table A.6 show all the parameters that need to be specified to execute `get_VarGradIG_from_TFDS.py` successfully.

Hyperparam	JSON variable name	Notes
m	`IG_m_steps`	Number of steps to approximate Integrated Gradient (IG)
x'	`IG_baseline`	Baseline image for IG. Available: "black" for a simple black image and "noise" for an image filled with Gaussian noise ($\mu = 0,\ \sigma = 1$)
n	`VarGrad_n_samples`	Number of noisy images to compute VarGrad (VG)

Table A.6: Parameters to compute score maps.

Appendix B

General remarks

This appendix contains general remarks relevant for this work, like a small explanation of the fluorescent markers used for the multiplexed protein map (MPM) protocol.

B.1 Indirect immunofluorescence markers description

In order to capture the distribution and amount of proteins inside a cell nucleus, the multiplexed protein map (MPM) protocol use a set of fluorescent markers called Indirect immunofluorescence (IF). Table B.1 shows a description of the most relevant markers. The identifiers and ids corresponding to the markers described in table B.1 can be consulted in table A.3.

Some of the markers used in the MPM protocol are strongly related with the *RNA polymerase* enzyme[1]. As it was explained in section 2.1 (see figure 2.3), the RNA polymerase it the enzyme responsible for starting the transcription process of genes (i.e., copping a sequence from a section of the DNA into a pre-messenger RNA (pre-mRNA) strand).

Marker name	Description
DAPI	*4',6-Diamidino-2-Phenylindole*, or DAPI, is a fluorescent stain that binds strongly to adenine–thymine-rich regions in DNA [Kap95]
GTF2B	*Transcription factor II B*, or TFIIB (also known as GTF2B), is an antibody that binds to the general transcription factor involved in the formation of the RNA polymerase II preinitiation complex [Lew04]

[1] An enzyme is a proteins that act as biological catalysts to accelerate chemical reactions.

SRRM2	*Serine/arginine repetitive matrix protein 2*, or SRRM2, is an antibody that binds to the protein that in humans is encoded by the SRRM2 gene and which is required for pre-mRNA splicing as component of the spliceosome. Along with the protein SON, SRRM2 is essential for Nuclear Speckles (NS)[2]formation [Ili+20]
SON	SON is protein that in humans is encoded by the SON gene. The protein binds to RNA and promotes pre-mRNA splicing, particularly of transcripts with poor splice sites. Along with the protein SRRM2, SON is essential for NS formation [Ili+20]
SP100	*SP100 nuclear antigen*[3], or SP100, is a gene that encodes a subnuclear organelle and major component of the PML (promyelocytic leukemia)-SP100 nuclear bodies [NCB]
PML	*Promyelocytic Leukemia*, or PML, is a protein encoded by the PML gene. PML is a nuclear body involved in oncogenesis (tumor suppressor) and viral infection. This subnuclear domain has been reported to be rich in RNA and a site of nascent RNA synthesis, implicating its direct involvement in the regulation of gene expression [BHBJ00]
PCNA	*Proliferating Cell Nuclear Antigen*, or PCNA, is a DNA clamp that acts as a processivity factor for *DNA polymerase* δ[4]in eukaryotic cells and is essential for replication [KLW05]
NCL	*Nucleolin*, or NCL, is an antibody that binds to a protein that in humans is encoded by the NCL gene. The protein is involved in the synthesis and maturation of ribosomes. It is located mainly in dense fibrillar regions of the nucleolus [ERA+88]
POL2RA_pS2	*RNA Polymerase II Phosphospecific (Ser2)*, or POL2RA_pS2, is an antibody that binds to the largest subunit of the RNA polymerase II (which is the enzyme responsible for transcribing DNA into pre-mRNA) [Nov]
CDK9	*Cyclin-dependent kinase 9*, or CDK9, is a protein encoded by the CDK9 gene and is involved in the regulation of transcription. CDK9 is a member of the cyclin-dependent kinase (CDK) family, which includes two main subgroups of kinases, those that mainly regulate cell cycle progression (including CDK1, CDK2, and CDK4/6) and those that control transcriptional processes (including CDK7, CDK8, CDK9, CDK12, and CDK13) [Cas+20]

CDK9_pT186	*Cyclin Dependent Kinase 9 Phospho-Thr186 Antibody*, or CDK9_pT186, is a molecule derived from human CDK9 around the phosphorylation site of T186 [BIO]
RB1_pS807_S811	*Retinoblastoma Protein pS807/pS811 Antibody*, or RB1_pS807_S811. Retinoblastoma Protein (RB1 or just RB) is a tumor suppressor protein, which prevents excessive cell growth by inhibiting cell cycle progression until the cell is ready to divide. [MB84]
PABPN1	*Polyadenylate-Binding Nuclear Protein 1*, or PABPN1 (also known as PABP-2), is a protein encoded by the PABPN1 gene, which is involved in the addition of a Poly-A tail to the pre-mRNA during the splicing process (see figure 2.4 on section 2.1.1) [MDW15]
SETD1A	*SET Domain Containing 1A, Histone Lysine Methyltransferase*, or SETD1A. The protein encoded by this gene is a component of a histone methyltransferase (HMT) complex that produces mono-, di-, and trimethylated histone H3 at Lys4. Trimethylation of histone H3 at lysine 4 (H3K4me3) is a chromatin modification known to generally mark the transcription start sites of active genes [Genb]
COIL	*Coilin*, or COIL. The protein encoded by this gene is an integral component of Cajal bodies, which are nuclear suborganelles involved in the post-transcriptional modification of small nuclear and small nucleolar RNAs [Gena]
EU	*5-Ethynyl Uridine*, or EU, is a molecule that binds to newly transcribed RNA [JS08]. This means that EU can be used to detect RNA synthesis in cells and/or predict Transcription Rate (TR)

Table B.1: Indirect immunofluorescence markers description. The first column shows the markers name, the second the identifier used on the implementation (parameters file) and the third a brief description of it.

[2]The NS (also known as *Splicing speckles*) are structures inside the cell nucleus in which the pre-mRNA is transformed into a mature messenger RNA (mRNA) (see section 2.1.1) [SL11].

[3]An *antigen* is a molecule that triggers the formation of antibodies (by bounding to its specific antibody or B-cell antigen receptor) and can cause an immune response.

[4]DNA polymerase delta, or DNA Pol δ, is an enzyme complex found in eukaryotes that is involved in DNA replication and repair.

List of Figures

List of Tables

Acronyms

4i iterative indirect immunofluorescence imaging. 2, 30–32, 96

Adam Adaptive Moment Estimation. 13, 52

ANN Artificial Neural Network. 3, 5, 10–15, 17, 18, 21, 95

CNN Convolutional Neural Network. iv, 1–5, 10, 18, 20, 29, 34, 37, 39, 45, 53–56, 61, 63, 64, 73, 76–78, 88, 99

DNN Deep Neural Network. 22, 23

GD Gradient Descent. 13

IF Indirect immunofluorescence. 30–32, 91, 93, 96, 100

IG Integrated Gradient. iv, 4, 23–27, 57, 58, 67, 77, 79, 90, 96

LIME Local Interpretable Model-Agnostic Explanations. 4

MAE Mean Absolute Error. 45, 56, 63–65, 97

MCU multiplexed cell unit. 2, 3, 95

ML Machine Learning. 3, 10

MLP Multilayer Perceptron. 12, 14

MPM multiplexed protein map. iv, 2, 3, 29–34, 45, 77, 78, 91, 96

mRNA messenger RNA. 2, 6–9, 29, 61, 70, 73, 76, 78, 93, 95

MSE Mean Squared Error. 52, 56, 97

NS Nuclear Speckles. 8, 70, 73–75, 92, 93

pre-mRNA pre-messenger RNA. 7–9, 33, 50, 70, 71, 73, 76, 78, 91–93, 95

ReLU Rectified Linear Unit. 11, 95

ROAR RemOve And Retrain. 76, 79

SG SmoothGrad. 4, 26

SGD Stochastic Gradient Descent. 13

SVM Support Vector Machine. 31, 32

TFDS TensorFlow Dataset. 35, 36, 38, 39, 41, 47–51, 85, 86, 89, 99

TR Transcription Rate. iv, 1, 9–12, 23, 24, 29, 33, 36, 37, 39, 42, 45, 46, 56, 61–63, 66–69, 73–79, 93, 97, 98, 100

VG VarGrad. iv, 4, 23, 26, 57, 58, 67, 77, 90

VGIG VarGrad Integrated Gradient. 26, 27, 58, 96

Bibliography

[Ade+18] J. Adebayo, J. Gilmer, I. Goodfellow, and B. Kim. *Local Explanation Methods for Deep Neural Networks Lack Sensitivity to Parameter Values*. 2018. arXiv: 1810.03307 [cs.CV].

[Ade+20] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim. *Sanity Checks for Saliency Maps*. 2020. arXiv: 1810.03292 [cs.CV].

[Bae+10] D. Baehrens, T. Schroeter, S. Harmeling, M. Kawanabe, K. Hansen, and K.-R. Müller. "How to Explain Individual Classification Decisions". In: *Journal of Machine Learning Research* 11.61 (2010), pp. 1803–1831.

[Bao+18] X. Bao, X. Guo, M. Yin, M. Tariq, Y. Lai, S. Kanwal, J. Zhou, N. Li, Y. Lv, C. Pulido-Quetglas, et al. "Capturing the interactome of newly transcribed RNA". In: *Nature methods* 15.3 (2018), pp. 213–220.

[BS00] R. G. Bartle and D. R. Sherbert. *Introduction to real analysis*. Vol. 2. Wiley New York, 2000.

[Ber+15] J. M. Berg, J. L. Tymoczko, G. J. G. Jr., and L. Stryer. *Biochemistry*. English. W. H. Freeman, 2015. ISBN: 1464126100.

[Bin+16] A. Binder, G. Montavon, S. Bach, K. Müller, and W. Samek. "Layer-wise Relevance Propagation for Neural Networks with Local Renormalization Layers". In: *CoRR* abs/1604.00825 (2016). arXiv: 1604.00825.

[BIO] BIOZOL. *CDK9 pT186 antibody, marker description*. https://www.biozol.de/en/product/ABX-ABX327917. Online; accessed 2021-05-03.

[Bis06] C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

[BHBJ00] F.-M. Boisvert, M. J. Hendzel, and D. P. Bazett-Jones. "Promyelocytic leukemia (PML) nuclear bodies are protein structures that do not accumulate RNA". In: *The Journal of cell biology* 148.2 (2000), pp. 283–292.

[BJ] T. Brown and T. B. (Jnr). *Simple representation of the gene expression process*. https://www.atdbio.com/content/14/Transcription-Translation-and-Replication. Online; accessed 2021-03-29.

[BHS14] A. R. Buxbaum, G. Haimovich, and R. H. Singer. "In the right place at the right time: visualizing and understanding mRNA localization". In: *Nature Reviews Molecular Cell Biology* 16.2 (Dec. 2014), pp. 95–109.

[Car+06] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. "CellProfiler: image analysis software for identifying and quantifying cell phenotypes". In: *Genome Biology* 7.10 (Oct. 2006), R100. ISSN: 1474-760X.

[Cas+20] M. Cassandri, R. Fioravanti, S. Pomella, S. Valente, D. Rotili, G. Del Baldo, B. De Angelis, R. Rota, and A. Mai. "CDK9 as a Valuable Target in Cancer: From Natural Compounds Inhibitors to Current Treatment in Pediatric Soft Tissue Sarcomas". In: *Frontiers in Pharmacology* 11 (2020), p. 1230.

[Che+16] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie. "Gene expression inference with deep learning". In: *Bioinformatics* 32.12 (2016), pp. 1832–1839.

[Cho17] F. Chollet. *Xception: Deep Learning with Depthwise Separable Convolutions*. 2017. arXiv: 1610.02357 [cs.CV].

[Cus+20] T. T. Cushnie, B. Cushnie, J. Echeverría, W. Fowsantear, S. Thammawat, J. L. Dodgson, S. Law, and S. M. Clow. "Bioprospecting for antibacterial drugs: A multidisciplinary perspective on natural product source material, bioassay selection and avoidable pitfalls". In: *Pharmaceutical Research* 37.7 (2020), pp. 1–24.

[Cyb89] G. Cybenko. "Approximation by superpositions of a sigmoidal function". In: *Mathematics of control, signals and systems* 2.4 (1989), pp. 303–314.

[ERA+88] M. S. ERARD, P Belenguer, M Caizergues-Ferrer, A Pantaloni, and F Amalric. "A major nucleolar protein, nucleolin, induces chromatin decondensation by binding to histone Hl". In: *European journal of biochemistry* 175.3 (1988), pp. 525–530.

[Fun89] K.-I. Funahashi. "On the approximate realization of continuous mappings by neural networks". In: *Neural networks* 2.3 (1989), pp. 183–192.

[Gena] Genecards. *COIL marker description*. https://www.genecards.org/cgi-bin/carddisp.pl?gene=COIL. Online; accessed 2021-05-04.

[Genb] Genecards. *SETD1A marker description*. https://www.genecards.org/cgi-bin/carddisp.pl?gene=SETD1A. Online; accessed 2021-05-04.

[GBC16] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. http://www.deeplearningbook.org. MIT Press, 2016.

[GHP18] G. Gut, M. D. Herrmann, and L. Pelkmans. "Multiplexed protein maps link subcellular organization to cellular states". In: *Science* 361.6401 (2018). ISSN: 0036-8075. eprint: https://science.sciencemag.org/content/361/6401/eaar7042.full.pdf.

[He+15] K. He, X. Zhang, S. Ren, and J. Sun. *Deep Residual Learning for Image Recognition*. 2015. arXiv: 1512.03385 [cs.CV].

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. *Identity Mappings in Deep Residual Networks*. 2016. arXiv: 1603.05027 [cs.CV].

[Hoo+18] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim. "A benchmark for interpretability methods in deep neural networks". In: *arXiv preprint arXiv:1806.10758* (2018).

[HSW89] K. Hornik, M. Stinchcombe, and H. White. "Multilayer feedforward networks are universal approximators". In: *Neural networks* 2.5 (1989), pp. 359–366.

[Hub] *Huber and MSE loss plot*. https://en.wikipedia.org/wiki/Huber_loss. Online; accessed 2021-04-26.

[Ili+20] bibinitperiodI. A. Ilik, M. Malszycki, A. K. Lübke, C. Schade, D. Meierhofer, and T. Aktaş. "SON and SRRM2 are essential for nuclear speckle formation". In: *ELife* 9 (2020), e60579.

[IS15] S. Ioffe and C. Szegedy. "Batch normalization: Accelerating deep network training by reducing internal covariate shift". In: *International conference on machine learning*. PMLR. 2015, pp. 448–456.

[JS08] C. Y. Jao and A. Salic. "Exploring RNA transcription and turnover in vivo by using click chemistry". In: *Proceedings of the National Academy of Sciences* 105.41 (2008), pp. 15779–15784.

[JD+13] W. JD, B. TA, B. SP, G. AA, L. M, and L. RM. *Molecular Biology of the Gene*. English. Pearson, 2013. ISBN: 9780321762436.

[Kap95] J. Kapuscinski. "DAPI: a DNA-specific fluorescent probe". In: *Biotechnic & Histochemistry* 70.5 (1995), pp. 220–233.

[Ker99] J. Kerr. *Atlas of functional histology*. English. Mosby International, 1999. ISBN: 0723430721.

[KB14] D. P. Kingma and J. Ba. "Adam: A method for stochastic optimization". In: *arXiv preprint arXiv:1412.6980* (2014).

[KLW05] J. Kisielewska, P. Lu, and M. Whitaker. "GFP—PCNA as an S-phase marker in embryos during the first and subsequent cell cycles". In: *Biology of the Cell* 97.3 (2005), pp. 221–229.

[Kor+11] V. I. Korolchuk, S. Saiki, M. Lichtenberg, F. H. Siddiqi, E. A. Roberts, S. Imarisio, L. Jahreiss, S. Sarkar, M. Futter, F. M. Menzies, C. J. O'Kane, V. Deretic, and D. C. Rubinsztein. "Lysosomal positioning coordinates cellular nutrient responses". In: *Nature Cell Biology* 13 (Mar. 2011), pp. 453–460.

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012), pp. 1097–1105.

[KSH17] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "ImageNet classification with deep convolutional neural networks". In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[Lew04] B. Lewin. *Genes 8*. Pearson Prentice Hall Upper Saddle River, 2004.

[MDW15] L. Muniz, L. Davidson, and S. West. "Poly (A) polymerase and the nuclear poly (A) binding protein, PABPN1, coordinate the splicing and degradation of a subset of human pre-mRNAs". In: *Molecular and cellular biology* 35.13 (2015), pp. 2218–2230.

[MB84] A. L. Murphree and W. F. Benedict. "Retinoblastoma: clues to human oncogenesis". In: *Science* 223.4640 (1984), pp. 1028–1033.

[NCB] NCBI. *SP100 SP100 nuclear antigen, marker description*. https://www.ncbi.nlm.nih.gov/gene/6672. Online; accessed 2021-05-03.

[Nov] Novusbio. *RNA Polymerase 2 Phosphospecific (Ser2), marker description*. https://www.novusbio.com/products/rna-polymerase-ii-polr2a-antibody_nb100-1805. Online; accessed 2021-05-03.

[PO+13] J. E. Pérez-Ortín, D. A. Medina, S. Chávez, and J. Moreno. "What do you mean by transcription rate?" In: *BioEssays* 35.12 (2013), pp. 1056–1062. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bies.201300057.

[RSG16] M. T. Ribeiro, S. Singh, and C. Guestrin. "Model-agnostic interpretability of machine learning". In: *arXiv preprint arXiv:1606.05386* (2016).

[Rui] M. Ruiz. *Animal eukaryotic cell diagram*. https://en.wikipedia.org/wiki/Cell_(biology). Online; accessed 2021-03-29.

[Rus+15] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. "ImageNet Large Scale Visual Recognition Challenge". In: *International Journal of Computer Vision (IJCV)* 115.3 (2015), pp. 211–252.

[SSR21] H. Saleem, A. R. Shahid, and B. Raza. "Visual Interpretability in 3D Brain Tumor Segmentation Network". In: *Computers in Biology and Medicine* (2021), p. 104410.

[Sen+20] A. Senior, R. Evans, J. Jumper, J. Kirkpatrick, L. Sifre, T. Green, C. Qin, A. Žídek, A. Nelson, A. Bridgland, H. Penedones, S. Petersen, K. Simonyan, S. Crossan, P. Kohli, D. Jones, D. Silver, K. Kavukcuoglu, and D. Hassabis. "Improved protein structure prediction using potentials from deep learning". In: *Nature* 577 (Jan. 2020), pp. 1–5.

[Seo+18] J. Seo, J. Choe, J. Koo, S. Jeon, B. Kim, and T. Jeon. "Noise-adding Methods of Saliency Map as Series of Higher Order Partial Derivative". In: *CoRR* abs/1806.03000 (2018). arXiv: 1806.03000.

[Shr+16] A. Shrikumar, P. Greenside, A. Shcherbina, and A. Kundaje. "Not Just a Black Box: Learning Important Features Through Propagating Activation Differences". In: *CoRR* abs/1605.01713 (2016). arXiv: 1605.01713.

[SSP+03] P. Y. Simard, D. Steinkraus, J. C. Platt, et al. "Best practices for convolutional neural networks applied to visual document analysis." In: *Icdar*. Vol. 3. 2003. Citeseer. 2003.

[SVZ13] K. Simonyan, A. Vedaldi, and A. Zisserman. "Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps." In: *CoRR* abs/1312.6034 (2013).

[Smi+17] D. Smilkov, N. Thorat, B. Kim, F. B. Viégas, and M. Wattenberg. "SmoothGrad: removing noise by adding noise". In: *CoRR* abs/1706.03825 (2017). arXiv: 1706.03825.

[Sni+12] B. Snijder, R. Sacher, P. Rämö, P. Liberali, K. Mench, N. Wolfrum, L. Burleigh, C. C. Scott, M. H. Verheije, J. Mercer, et al. "Single-cell analysis of population context advances RNAi screening at multiple levels". In: *Molecular systems biology* 8.1 (2012), p. 579.

[SL11] D. L. Spector and A. I. Lamond. "Nuclear speckles". In: *Cold Spring Harbor perspectives in biology* 3.2 (2011), a000646.

[Spr+14] J. Springenberg, A. Dosovitskiy, T. Brox, and M. Riedmiller. "Striving for Simplicity: The All Convolutional Net". In: (Dec. 2014).

[SJ+60] R. G. Steel, H James, et al. *Principles and procedures of statistics: with special reference to the biological sciences*. Tech. rep. 1960.

[SF19] B. van Steensel and E. E. Furlong. "The role of transcription in shaping the spatial organization of the genome". In: *Nature Reviews Molecular Cell Biology* 20.6 (2019), pp. 327–337.

[SLL20] P. Sturmfels, S. Lundberg, and S.-I. Lee. "Visualizing the Impact of Feature Attribution Baselines". In: *Distill* (2020). https://distill.pub/2020/attribution-baselines.

[STY17] M. Sundararajan, A. Taly, and Q. Yan. *Axiomatic Attribution for Deep Networks*. 2017. arXiv: 1703.01365 [cs.LG].

[Vil] M. R. Villarreal. *Transcription steps*. https://www.ck12.org/biology/transcription/lesson/Transcription-of-DNA-to-RNA-BIO/. Online; accessed 2021-03-29.

[Vog+10] C. Vogel, R. de Sousa Abreu, D. Ko, S.-Y. Le, B. A. Shapiro, S. C. Burns, D. Sandhu, D. R. Boutz, E. M. Marcotte, and L. O. Penalva. "Sequence signatures and mRNA concentration can explain two-thirds of protein abundance variation in a human cell line". In: *Molecular systems biology* 6.1 (2010), p. 400.

[Wan+93] D. G. Wansink, W. Schul, I. Van Der Kraan, B. Van Steensel, R. Van Driel, and L. De Jong. "Fluorescent labeling of nascent RNA reveals transcription by RNA polymerase II in domains scattered throughout the nucleus". In: *The Journal of cell biology* 122.2 (1993), pp. 283–293.

[Wik21] Wikipedia. *Primary transcript — Wikipedia, The Free Encyclopedia*. http://en.wikipedia.org/w/index.php?title=Primary%20transcript&oldid=1005412823. [Online; accessed 29-March-2021]. 2021.

[WL11] C. L. Will and R. Lührmann. "Spliceosome structure and function". In: *Cold Spring Harbor perspectives in biology* 3.7 (2011), a003707.

[You06] R. M. Youngson. *Collins dictionary of human biology*. Collins Publishers, 2006.

[ZF14] M. D. Zeiler and R. Fergus. "Visualizing and understanding convolutional networks". In: *European conference on computer vision*. Springer. 2014, pp. 818–833.

[Zha+88] W. Zhang et al. "Shift-invariant pattern recognition neural network and its optical architecture". In: *Proceedings of annual conference of the Japan Society of Applied Physics*. 1988.