LU-TP 22-43
June 2022

Cancer Driver Gene Detection using Deep Convolutional Neural Networks on H3K4me3 Enrichment Profiles

Marc Pielies Avellí

Department of Astronomy and Theoretical Physics, Lund University

Master thesis (FYTM03, 30 ECTS)
Supervised by Dr. Victor Olariu and Dr. Mattias Ohlsson

LUND
UNIVERSITY

Abstract

In spite of our current knowledge regarding the biology underlying cancer genesis, reliable methods for the discovery of cancer driver (CD) genes are still in great need. The rather recent incorporation of epigenetic markers to the cancer paradigm has nevertheless opened the door for the development of new computational approaches to the problem.

This work is aimed to study the enrichment of certain genome regions with the histone post-translational modification (PTM) H3K4me3. This epigenetic marker can be used to distinguish cancer driver genes from neutral genes (NGs). To this end, a convolutional neural network (CNN) comparing H3K4me3 enrichment profiles for matching healthy and cancer samples is proposed and evaluated. The obtained results for OriGENE, the presented model, show promise in pan-cancer but also tissue-specific cancer driver gene detection.

Popular introduction to the project: On the miracle of cancer therapies, the genome bible and the power of teachers

According to the World Health Organization (WHO), in 2020 nearly 10 million people died of cancer worldwide. Cancers are known to emerge when groups of cells divide and grow at higher rates than usual. Therefore, if one wants to understand the Genesis of the disease, it is compulsory to discover the core elements regulating the aforementioned processes.

These fundamental elements are the genes, discrete pieces of DNA encoding the information about "how to create and maintain alive" almost any living being that one can imagine. This information is encoded using four basic molecules that are named with the letters A,T,G and C. Our genes can then be thought as the words constituting a book that could be called The Human Genome. Since the book is genuinely large and has had a huge impact on our understanding of human nature, we'll baptize it the Genome Bible.

Let's assume that our body works like the Theology faculty at a certain university. Even though the text book in both cases is the same in all the classes (which in the analogy would be the cells), the parts of the book that a professor is going to read and the sections that will be skipped are at the end as important as the actual text. Moreover, not only what is read matters but also how the text is read plays probably one of the most important roles: the interpretation and enthusiasm with which the theology professor transmits the message to the students will deeply shape their minds and their future behavior. These factors, which are not intrinsic to a message that has only suffered slight mutations over the past centuries, have led to outcomes as different as the Crusades and the creation of some of the most altruistic and useful NGOs.

The same picture arises in the field of genetics, but in this field it turns out to be even more explicitly a matter of life and death. Enthusiasm, interpretation and book chapter selection are in our case mapped into gene expression levels, gene roles within tissues and gene silencing and activation. All these features and processes are regulated by the so-called epigenetic factors, external elements to the DNA that control how the encoded information is expressed and read.

In our work, we will try to pinpoint which genes are prone to be related with cancer genesis. To do so, a specific type of epigenetic factors that induce changes in the proteins around which the DNA is wrapped will be studied. Our focus will therefore be on the behavior of the teacher and the lecture itself rather than the book. The characterization of the genes will be done by means of machine learning related techniques. This will allow us to find the patterns in the epigenetic factors that lead genes to be expressed in aberrant manners. The results obtained in the project will shed light on the role of epigenetics in tumorigenesis and could have implications in drug and therapy development.

Contents

1 Introduction

Once one starts to delve into the world of biology, an honest truth slowly reveals itself: life is an extremely fragile process. This fragility is manifested at its best at the cellular level, where the perfect trade-off established between cell death and proliferation continuously pushes our tissues away from a fatal fate.

Even though there are many factors that can alter the trade-off between cell death and division, one can most of the times track down the origin of the complications to the gene level. More precisely, problems tend to stem from the network of genes that are active and silent given a specific cellular context. Misregulation or changes in the functionality of a certain set of genes involved in cell growth and division, for example, will lead the cells to grow and multiply out of control. This process, outlined in Figure 1, is known as tumorigenesis, carcinogenesis or oncogenesis, and is the ultimate cause of cancer.

Hence, it is of utmost importance to pinpoint which genes can be linked with the emergence and development of the disease. These genes are called cancer drivers (CDs), and can be further divided into two categories: oncogenes (OGs), the upregulation of which is directly linked to carcinogenesis, and Tumor Suppressor Genes (TSGs), which inhibit the aberrant expression of the first group. For classification purposes, oncogenes and tumor suppressors are usually compared with housekeeping genes. These are highly conserved neutral genes (NGs) required for the proper functioning of any kind of cell, independently of the cell's identity.

Originally, the scientific paradigm regarding the genetic nature of cancer contemplated mainly the mutations that happened in the bodies of the genes. Such changes could allow the proteins encoded in OGs to acquire undesired functions, for instance. They could also reduce or even abolish the expression of TSGs. The first computational methods were therefore developed trying to tackle the problem from a purely genomic and mutational perspective [1]-[10].

However, it became more and more apparent that higher than average mutation rates could not explain all the known cancer drivers [11], [12]. This realization led part of the scientific community to shift their attention from the DNA and the genes to the external elements that regulate how DNA is actually read and expressed. Although there are many definitions, the study of "heritable changes in gene function that do not entail a change in DNA sequence" is known as epigenetics [13].

The most studied epigenetic markers as of today are probably histone post-translational modifications (PTMs) [14]. Histones, the proteins around which DNA is coiled, can undergo a number of modifications after they are translated. In this work we will focus on methylations (monomethylation me1, trimethylation me3) and acetylations (ac) of the lysines (K) found at the tails of such proteins, in particular the last aminoacids of histone H3. The methyl and acetyl groups added can display for example a harbouring effect for transcription factors (H3K4me3), leading to a higher transcription of the surrounding

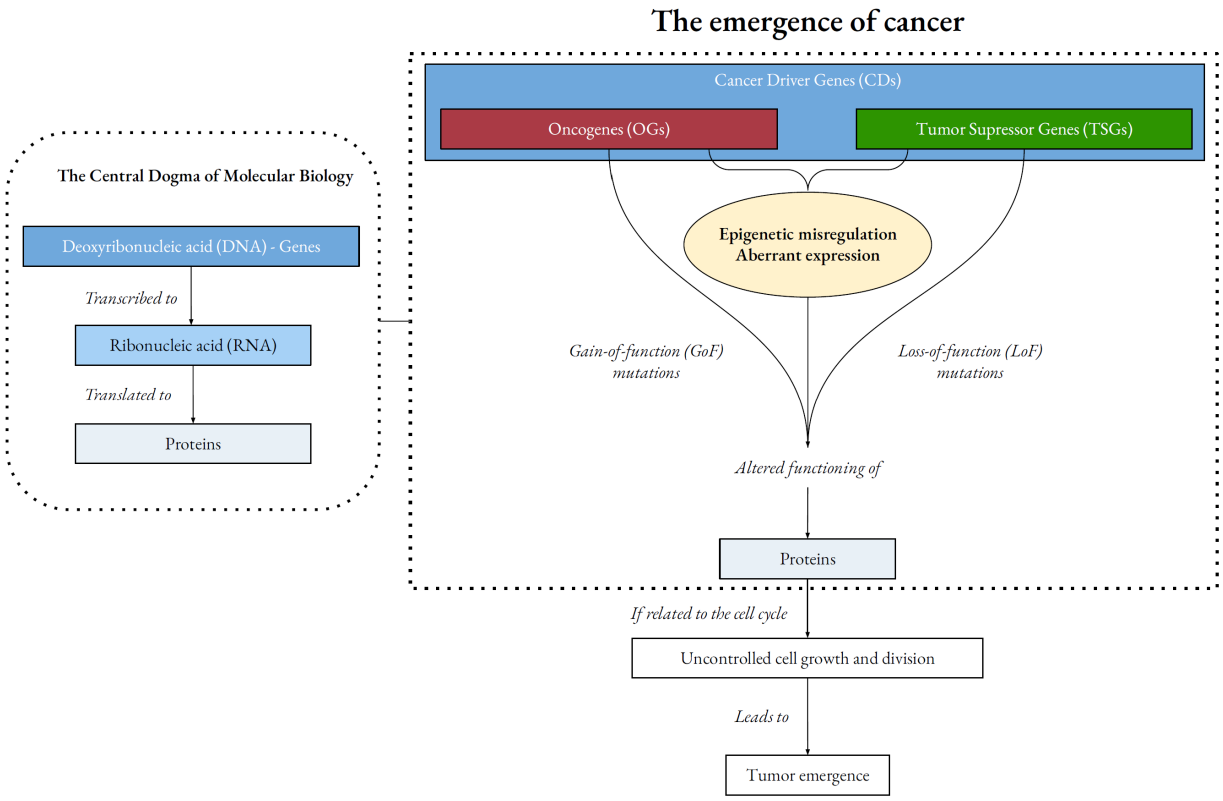


Figure 1: **Cancer emergence schema.** On the left, the central dogma of molecular biology is presented. Genes are first transcribed into RNA. RNA will then be translated into proteins with a role, e.g. functional or structural. On the right, the emergence of cancer is outlined from the genetic level to a macroscopical scale. Apart from the well studied gene mutations, elements such as misregulation of epigenetic markers can also be linked with cancer development. Note that the path leading to the macroscopical tumor is not unique.

DNA [15]. They can also have a silencing nature (H3K27me3) or point towards other functional domains in the genome such as enhancers and superenhancers (H3K4me1 and H3K27ac). The hypothesis that different markers and combinations of them can encode functional information beyond the genetic material is known as the histone code [16].

The spatial distribution in the genome of such modifiers can in fact hide information linking certain genes with cancer. Broad domains of H3K4me3, for example, have been shown to be an intrinsic property of tumor suppressor genes [17], while repressive markers such as H3K27me3 can in turn silence oncogenes [18].

The power of epigenetic markers in functional classification tasks has also already been tested. In the Roadmap epigenomics project [19], a Hidden Markov Model (HMM) using several markers was trained pursuing to find different functional domains. Jie Lyu et al. developed a machine learning approach that involved epigenetic features such as peak width (how long is the region where we can find the marker) and height (how significant is the concentration of the marker locally), and combined them with genomics data in order to discover OGs and TSGs [11]. Another notable project that merged machine learning and cancer gene prediction was conducted by R. Schulte-Sasse et al. [12]. The group also integrated many data types including DNA-methylation and trained an explainable computational method based on graph convolutional networks.

Despite all these advances in the development of computational tools for cancer gene prediction, we found the following fronts to be poorly explored. We noticed that epigenetic information had played an auxiliary role in most of the research projects published to date. Bo Xia et al. already started to dive deeper into feature extraction from such markers in a cell identity gene (CIG) identification task [20]. However, the epigenetic features to analyze had been manually curated and introduced to a neural network later in all mentioned studies. Furthermore, tissue- and patient-specific algorithms were still in great need, since most projects adopted a pan-cancer perspective. The possibility to use machine-learning to extract information from the raw epigenetic data, and the ability to use it in tissue- and patient-specific contexts constituted still a gap in our current knowledge.

The goal of this project was therefore to develop a new approach to the cancer driver gene classification problem addressing these needs. To this end, we explored the ability of deep convolutional neural networks (CNNs) to retrieve information from the most basic representation of the epigenetic marker signals, their enrichment profiles. From H3K4me3 profiles for single genes, the networks would bring out and integrate the necessary features to classify each gene as CD or NG. The designed CNN would also compare, for liver and lung cancer patients, the differences gene-to-gene between tumor and healthy samples.

Our work, which culminated with the creation of the deep convolutional neural network OriGENE, is presented in this thesis as it follows. Sections 1 and 2 introduce the studied problem and motivate the project. Section 3 provides the reader with the necessary theoretical background. The obtained results and subsequent discussion are presented in Section 4. The main findings of this work are then summarized in the Conclusions.

2 Problem description

In this thesis we propose a machine learning model aimed to classify genes as either Cancer Drivers (CDs) or Neutral Genes (NGs). The model is tested in several case scenarios.

This section will cover what data the model requires and how it can be exploited for the classification problem. First, the used data will be introduced and described. Subsequently, an outline of how the algorithm analyzes the data will be presented. Finally, we will show why the proposed approach improves our current knowledge.

2.1 The data

The genes. The first task in a gene classification problem is to find genes that can fall into the desired categories or classes. These genes need to have well known ground truth labels, which will define their target classes. Genes labelled as CDs (both OGs and TSGs) will be assigned the target class 1, while NGs will be assigned the target class 0.

The number of well established cancer driver genes was still scarce by the time this thesis was written. This fact deeply impacted our project and is discussed in the following sections. The original gene set was constituted by 480 CDs and 489 NGs. Around 80% of these genes were stored in a combined Training and Validation set, while the remaining part was put aside for testing the final model.

For the pan-cancer stage of our study, the curated list of genes that are known to play a role in cancer (CDs) and the opposite set of housekeeping genes (NGs) were obtained from [11][1]. Their original data sources were the Cancer Gene Census (CGC v.87) from COSMIC [21] for CDs, and T. Davoli et. al. [5] for NGs. Cancermine [22] curated genes substituted the original cancer drivers when the project was brought to a tissue-specific level.

The samples. The findings presented in this thesis concern the dataset GSE67471. The data was provided by [17] and found at the Gene Expression Omnibus [23]. The said dataset contains data from four cancer patients. Two of them were diagnosed with liver cancer (Liver I, II) while the other two were known to have lung cancer (Lung I, II). Samples from the cancer tissue of interest and the respective matching healthy tissue were extracted for every single patient.

More information regarding data accessibility can be found in Appendix A.

The sequences. The epigenetic marker of interest was the addition of three methyl groups (trimethylation) to the fourth-to-last amino-acid in histone H3 (H3K4me3). Figure 2 shows the levels of this marker along an example gene. We will refer to these plots as H3K4me3 enrichment profiles for specific genes. H3K4me3 enrichment profiles describe how the relative amount of trimethylation, represented in the y-axis, changes depending on the position in the genome, represented in the x-axis. These profiles were retrieved for

[1] Chromosome Y CDs were skipped in order to generalize to any sample, independently of sex.

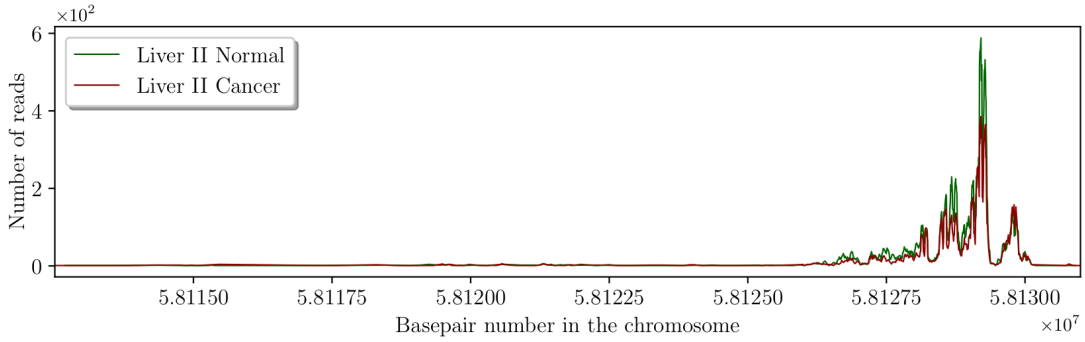


Figure 2: **H3K4me3 enrichment profiles for the gene C16ORF80**. C16ORF80 is neutral gene (NG) encoded in the negative strand. The shown profiles correspond to the healthy sample, plotted in green, and the cancer sample, plotted in red, for the patient Liver II. Note that H3K4me3 enrichment profiles are highly localized around a short region of the gene.

all curated genes from the eight tissue samples. The profiles from healthy and matching cancer samples for each gene were then paired, and constituted the input data to our model.

Note the critical difference between cancer sample and cancer driver gene. Cancer samples will be used as inputs, together with their healthy counterparts. Cancer driver (CD) is in turn a label that refers to the target class of a certain gene, not a sample. Hence, this label does not concern the origin or health status of the input tissue.

2.2 Motivation

Recent attempts to use machine learning techniques in gene category classification problems introduced to the networks already curated sets of genetic or/and epigenetic features as inputs [11],[20]. On the contrary, we wanted to follow the late trend to work as much as possible with the raw original data, and put to trial "the machine learning dogma".

We believed that a CNN should *a priori* have the capability to extract the necessary information from the unprocessed H3K4me3 profiles. To perform at its best the classification task, the network would create a set of tailored filters bringing out different signal features. All the extracted features would then be integrated in the final set of fully connected layers. The network would lastly output a probability for each gene to be a cancer driver, with the genes being described in our model by the pairs of profiles used as inputs.

Thus, the most fundamental novelty of our work was the use of a CNN to perform an unconstrained, multi-scale and spatially resolved feature extraction process from the input signals.

This approach had some drawbacks. For instance, the interpretation of the elements that the network could be looking at was not straightforward. Moreover, given the considerable length of the genes, a significant size reduction of the sequences was required in order to keep the number of parameters of the model low.

Nevertheless, the approach had definitely its advantages, presented below.

This method could study the changes that H3K4me3 profiles for each gene undergo when a patient develops a cancer, both for neutral genes and cancer drivers. The architecture would end up naturally analyzing simple features as peak width and height. These features could be compared between CDs and NGs, but also between both samples if relevant. More importantly, the network could go further in the abstraction and complexity of the studied features.

In addition, this approach allowed the project to be extended to a cancer-specific level.

Finally, the data requirements of the method were relatively low. In our multi-omics era, this project relied only on data for some epigenetic markers in single patients. The network could still be generalized and include input tracks with genomic information if needeed[2].

As a summary, all points above motivated the present thesis, aimed to tackle the cancer gene classification problem. This was done by means of a CNN, which could compare and analyze the enrichment profiles of the epigenetic marker H3K4me3 for single genes.

[2]Variant calling to find genetic mutations could be coupled to the epigenetic information already in use.

3 Theoretical Background

This section will cover the most important concepts and tools required to contextualize our work. Both basic concepts and technical details regarding the machine learning side of the project can be found in Appendix B.1.

3.1 Biological insights

The most intuitive way to get an understanding of the system we studied is to start from the fundamental building blocks and zoom out little by little, in a bottom-up approach.

The first building block is Deoxyribonucleic acid (DNA). DNA is a macromolecule constituted by a set of paired nucleotides matching the nucleobases Adenine with Thymine and Guanine with Cytosine. Paired nucleotides are separated 3.4 Å from each other, and are then coiled in a double helix shape of around 20 Å of diameter[3].

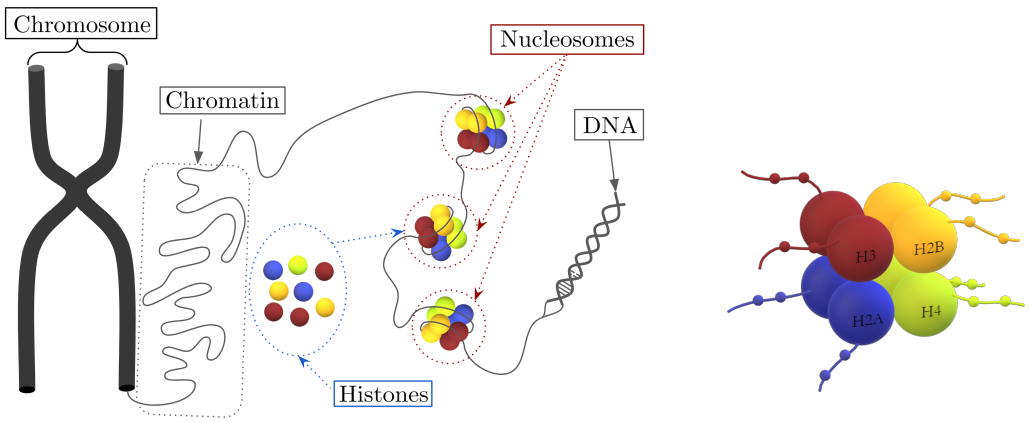
The second group of building blocks, which are the most relevant for the project, are the histones. Histones are proteins with a high content of the aminoacids arginine (R) and lysine (K). These proteins play a crucial role in DNA packaging, providing structural support to the chromosomes, but are also involved in the regulation of gene expression. Histones H2A, H2B, H3 and H4 are doubled and assembled into a complex called histone octamer, which is rather cylindrical, illustrated in Figure 3.

DNA is wrapped approximately 1.7 times around the octamer, which corresponds to 147 basepairs (bp). The histone octamer with the DNA wrapped around is called a nucleosome core. Then, histone H1 acts as a linker histone. Finally the nucleosome core, together with the linker DNA and H1 constitute the so-called nucleosome. In its chromatin state, DNA is distributed as sets of nucleosomes, which have a direct impact on DNA accessibility and packaging [24].

Histones, and especially the last positively charged aminoacids at their flexible N-terminal, can be covalently modified after being translated. Some of the most common modifications that these tails flanking the ends of the central histone fold can undergo are phosphorilation, acetylation, mono- or pluri- methylation and ubiquitination. As mentioned in the sections above, this work will be devoted to the study of H3K4me3, i.e. the trimethylation of the fourth lysine (K4) of the histone H3.

The addition of methyl or acetyl groups to different aminoacids of the histones has been shown to play a signaling role, by recruiting other proteins with specific domains able to recognize such markers [25], for instance transcription factors (TFs). These molecular markers and combinations of them can hence encode inheritable functional information [16], as the control on the regulation and expression of the surrounding genes that they exert can be maintained after cell division [26].

[3]For context, we remind the reader that the hydrogen atom has a diameter of ~ 1Å.



(a) (b)

Figure 3: **Global picture: from the double helix to the chromosome.** a) The DNA double helix is wrapped ~ 1.7 times around the histone octamer, originating the most basic DNA packaging unit: the nucleosome. Nucleosomes, which are connected by linker DNA in this image, can be condensed even further to create the chromatin fiber that will constitute the chromosomes. b) Histones H2B, H2A, H3 and H4 are doubled in the histone octamer. The last aminoacids of the said structural proteins, illustrated as small beads in their flexible tails, will undergo post-translational modifications that will display a functional signaling role, shaping the accessibility of other proteins to the DNA.

Promoter regions, enhancers, silencing domains and gene bodies, for example, constitute some of the main functional domains in the genome. These domains can be characterized by the presence of certain histone modifiers.

Promoters are specific sequences in the DNA that define where and in which direction will the RNA-polymerase start the transcription. Promoter regions tend to be located immediately upstream the transcription start site (TSS) and are marked with a significant enrichment of H3K4me3 when the genes are active [27].

Enhancers are DNA regions where proteins will be gathered, increasing the transcription of closeby genes[4]. Enhancers are open chromatin domains, with no nucleosomes, which appear as H3K4me3 devoided regions. This depletion is accompanied by an increase of the competing enhancer marker H3K4me1, while H3K27ac is also found in enhancer and superenhancer domains.

Histone PTMs can also display the opposite effect in silencing domains. Important repressive markers such as H3K9me3 and H3K27me3, found in tighly packed DNA regions known as heterochromatin, contribute to the downregulation of nearby genes [18][28]. Other markers such as H3K36me3 point towards the body of the genes and have been proven to be involved in DNA repair and stability.

As a summary of this section, DNA is coiled around a rather cylindrical complex of proteins, the so-called histones, creating a fundamental DNA packaging structure: the nucleosome. These proteins can accept signaling molecules at their flexible tails that will characterize the functional role of the surrounding DNA. Higher structures like chromatin fiber or the chromosomes are out of the scope of this project.

3.2 Retrieving the enrichment profiles of the epigenetic marker H3K4me3

This section will guide the reader through the data acquisition process, pursuing to provide an overview of the origin, appearance and content of the signals we will work with.

Chromatin ImmunoPrecipitation (ChIP). A technique called Chromatin Immuno-Precipitation (ChIP) can be used to determine the level of enrichment of distinct DNA regions with various histone PTMs.

The proteins of interest, in our case the histones, are initially cross-linked with the DNA material from numerous nuclei using formaldehyde[5]. The cross-linked DNA is subsequently sheared into mononucleosomes, which will have small DNA fragments attached. For this purpose, the enzyme Micrococcal nuclease (MNase) is used to cut the DNA region linking nucleosomes and digest the free DNA ends toward the nucleosome. Since MNase cannot go

[4]The regulated genes are not necessarily close in the DNA chain as the 3D folded structure can lead to interactions between distant regions.

[5]In the protocol $\sim 5 \cdot 10^7$ nuclei were yielded [17].

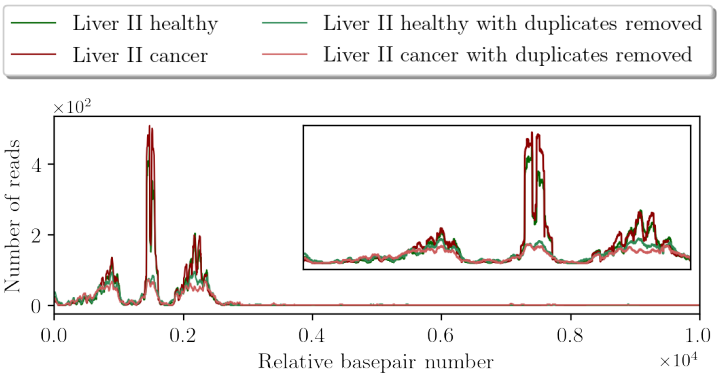


Figure 4: **Final H3K4me3 profiles illustrated**. The gene CASP9, a tumor suppressor gene (TSG) encoded in the negative strand is shown. A well conserved nucleosome location can be seen. The wiggles with nucleosome positioning information are still partially kept when removing duplicates, and would be missed in a broad domain analysis. Note that the profiles will be flipped and cropped if necessary in the data preprocessing steps.

further than the nucleosome, the fragments left will implicitly contain information about the expected location of such octamers in the genome [29].

Seeking to separate H3K4me3-bound DNA fragments from the rest, the described DNA-protein complexes are immunoprecipitated with protein-specific antibodies attached to magnetic beads. Considering that the DNA fragments are still linked with the proteins, cross-linking must be reversed. The remaining pieces of DNA are purified and amplified using the widespread Polymerase Chain Reaction (PCR) technique[6].

Sequencing. The nucleic acid sequences of the purified DNA fragments are at this point obtained with the next generation sequencing technology. The sequencing machine, in our study, outputs files containing the first 50 basepairs of each single stranded fragment. These short sequences, which can come from any of the two strands, are called single-end reads. The reads will then be aligned and mapped to a reference human genome assembly called hg38. This mapping will allow us to track the original location of each read.

Enrichment profile obtention. Once the reads have been mapped to specific genome regions, one can pile them up and create a bin-like enrichment profile. The enrichment intensity for the desired epigenetic marker at every single basepair will correspond to the number of reads that overlap in that particular genome location[7]. An example of the enrichment profiles for H3K4me3 was already presented as Figure 2 in Section 2.

After normalization, standardization and relative alignment of the different samples, the signals will be ready to be analyzed. An example of the final signals is shown in Figure 4. The described procedure is further explained in Appendix B.2.

[6]10 PCR cycles were performed in [17].

[7]This step is illustrated in Figure 16 at the Appendix C.

As it can be noticed in Figure 4, one can expect to find a high enrichment of H3K4me3 at the beginning of the gene, around its promoter and TSS. The profile can also be extended towards the gene body.

One can also start to identify H3K4me3 devoided sections or gaps related to enhancer regions. In addition, it can be seen that the signal presents oscillations at many different spatial frequencies, which will be shown to contain both functional and structural information.

Finally, the coupled peaks and pillars, which are smoothed but kept when removing exactly duplicated reads, point implicitly towards the location of the nucleosomes (For more information, see Appendix C).

4 Results and Discussion

The main section of this thesis is divided in two distinct blocks. The first part will cover the qualitative pieces of information and preliminary tests that would motivate and shape the future development of our deep learning approach to the problem. In the last section, the details of our proposed CNN are presented and its performance in different contexts is evaluated and discussed.

4.1 Preliminary considerations

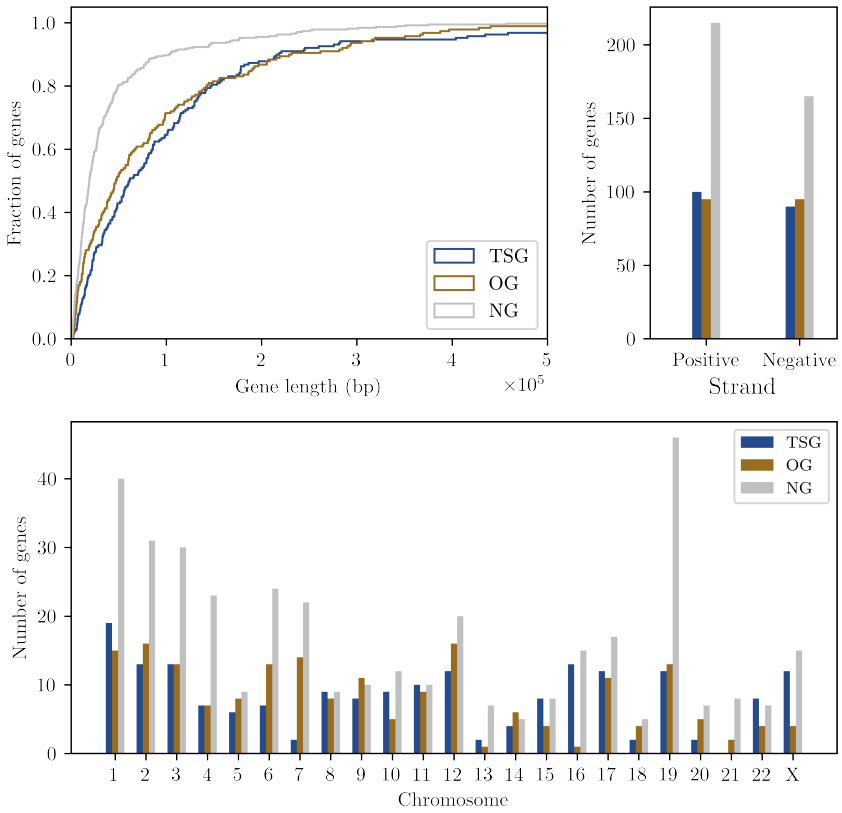


Figure 5: **General overview of the gene set**. Top left: cumulative gene length distribution. Top right: strand distribution. Bottom: Chromosome distribution.

The distribution of the chosen genes across chromosomes, strands and their lengths were studied before starting a detailed analysis of the profiles. The purpose of this step was to bring to light trivial imbalances between genes corresponding to different classes, if any.

The top left plot in Figure 5 shows that the chosen NGs are on average shorter than the curated CDs, including OGs and TSGs. More than 80% of the NGs already appear before 10^5 bp. However, one should keep in mind that traditional ways to detect OGs and TSGs have been based on gene mutations, especially in coding regions. This could introduce a clear bias towards the discovery of longer cancer drivers. We can therefore only say that the already known and verified cancer drivers are indeed slightly larger on average.

The top right plot in Figure 5 shows that both NGs and CDs are almost equally split between the two strands. The total number of NGs doubles TSGs and OGs for class balance purposes. As expected, there is no asymmetry or preference for CDs to be encoded in a specific strand.

The plot at the bottom of Figure 5 shows how CDs and NGs are spread over all the different chromosomes. From the chromosomal distribution one can see that no TSGs were included from chr 21 and a significant number of NGs come from chr 19. Some chromosomes, e.g. 8 and 9, present a high number of CDs when compared with their NGs.

None of the gathered pieces of information points towards a clear natural tendency to find CDs and NGs in distinct chromosomes, strands nor length ranges. Consequently, these elements will not be introduced in our model.

4.1.1 Structural and functional information

At the beginning of the project we considered histone PTMs, for instance H3K4me3 or H3K27ac, to play mainly a functional and signaling role. However, the structural information encoded in the profiles of these markers was found to be way more significant than what we expected originally, in agreement with [30].

Previous studies such as [17] and [18] had approached cancer driver gene characterization from a broad domain perspective, studying peaks that spanned more than 4 kbp. Nevertheless, the H3K4me3 signals for different tissues and samples obtained in independent experiments did actually show similar trends down to a $\sim$50 bp resolution. Figures 4, 6 and 7 illustrate this phenomenon. This information would be completely lost when approaching the problem in terms of broad domains and averages over different datasets, and could be potentially useful for our purposes.

Sequence independence. The spatial distribution of the epigenetic marker H3K4me3 constitutes a signature of the gene. This fact is revealed in Figure 6, where inter-gene variability (from column to column) is considerably more significant than the intra-gene feature differences (from row to row). As a result, sequences or profiles corresponding to a specific gene, even if they belong to different individuals or have a different health status, cannot be considered fully independent.

The lack of independence between sequences must be taken into account when designing a machine learning-based approach. More specifically, sequences for the same gene should not be put both in Training and Validation sets. Otherwise, the network would recognize

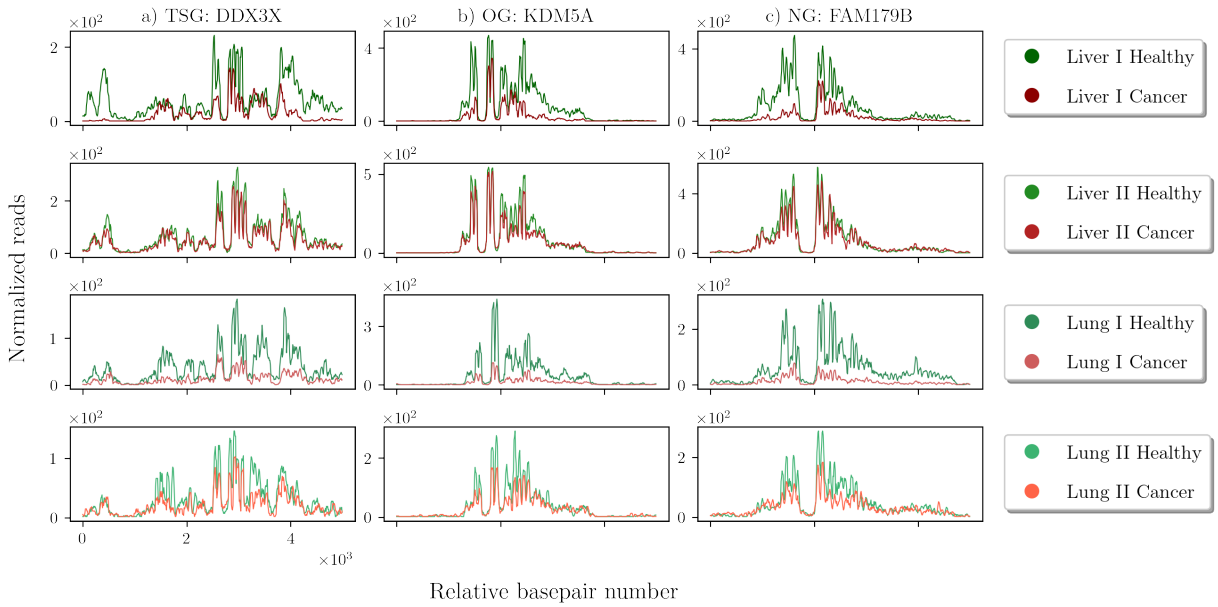


Figure 6: **Highly shared structure for the same gene in different samples**. Liver I, Liver II, Lung I and Lung II paired samples are shown in different rows for the following genes presented in columns a) TSG DDX3X, b) OG KDM5A and c) NG FAM179B. Signals were subsampled to a scale of 10 bp $\equiv$ 1 dbp to ease visualization. Each sequence starts around 2000 bp before the transcription start site (TSS) of the gene. This range ensures that the promoter region of the gene, found just before the TSS, will be captured.

the whole profile, especially for the healthy sequences, as something already seen during the training stage. This could lead to non-real astonishing validation performances.

Conversely, the performance of the model could drastically drop when assessing the actual capability of the network to predict genes from an independent test set with new genes. This would be an example of overfitting to the validation set, and constituted one of the strongest limitations of our approach. It meant that in our case introducing more samples would not necessarily provide a better model, but instead could risk its generalization power, unlike most machine learning related projects.

Our approach is thus limited by the available curated sets of OGs and TSGs and their reliability, especially if one wants to pursue a tissue-specific project. This results deeply shaped our future work.

Pattern translation. Figure 7 depicts some concepts that deeply motivate using CNNs when trying to solve the cancer gene classification problem.

The first concept, pattern or structure translation, is illustrated in Figure 7a). In the figure one can see that, once more, almost the same chromatin structure characterizes different samples and tissues. This similarity is most evident for the healthy samples, depicted in green. In this particular case, the arbitrarily chosen gene ZNF394 is found close to another one, ZKSCAN5, encoded in the opposite strand. This allows us to observe the enrichment at the promoter regions of both genes simultaneously.

As explained in the previous section, basic building blocks of the signals such as the enhancers, which appear as H3K4me3 devoided wells (highlighted in gold), and nucleosome-related structures like the coupled pillars (highlighted in blue) are shared among different genes. These elements can appear translated, permuted or even rotated if the strand information is not considered.

CNNs are known to display their maximum potential when identifying translated patterns, and hence they are especially suitable for this problem.

Aligned comparison. Figures 7b) and c) illustrate why it is interesting to keep the spatial information of the signals. The local and aligned juxtaposition of the tracks for matching healthy and cancer tissues allows a CNN to study the combined behaviour of both signals in a spatially resolved fashion.

With an input like b), CNNs have the capability to analyze how the methylation pattern is extended towards the gene body in TSGs. Such networks can also assess if the pattern is shortened or milder in the cancer track. These features could be linked with an insufficient expression of the gene in a cancer scenario, as presented in [17].

The exact opposite case arises in Figure 7c). The figure also shows how H3K4me3 can implicitly point towards aberrant gene expression levels. Here, the cancer sample's H3K4me3 signal exceeds its healthy counterpart. The higher enrichment band could be accompanied by an increased harbouring effect for transcription factors [15]. The previous process can lead to overexpression-induced oncogenesis, one of the types proposed in [12], which is not

necessarily linked to somatic mutations. This information, usually provided by RNA-seq data, could already be hidden partially in H3K4me3 signals.

In addition, the first nucleosome downstream of a start site has been shown to exhibit differential positioning in active and silent genes [30]. This implies that cancer-induced changes in the activity of a gene could also give rise to shifts in the profiles, which could be measured from the aligned inputs.

It must be stressed that these examples were arbitrarily picked for illustration purposes, with the only aim to convey the following message: the discussed features or other similar signal characteristics, if conserved and shared among sequences within a certain class, would cause the creation of specific sets of filters in the first set of layers of a deep convolutional neural network. Moreover, these filters would emerge as a natural consequence of the link weight optimization when training a neural network using back-propagation (See Appendix B.1). The outputs from such filters can then be reintroduced in the final steps of the algorithm, where the classifier decides whether the found patterns and the way in which they are related spatially are useful in order to predict the class correctly.

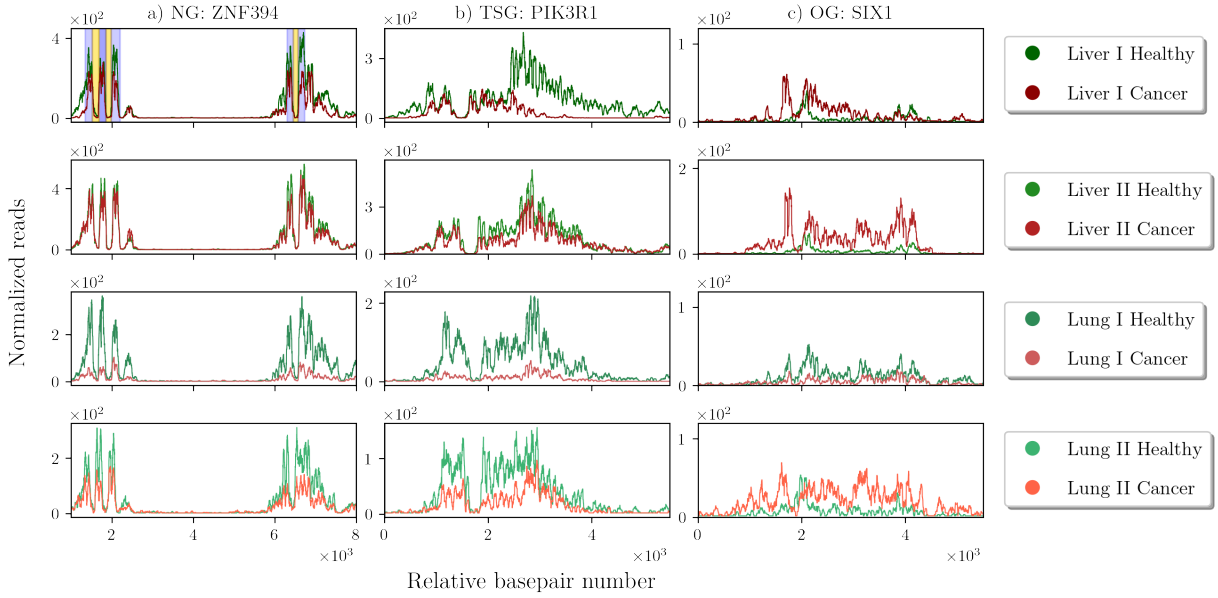


Figure 7: **Pattern translations and local feature comparisons motivate the use of CNNs.** a) Pattern translations. Enhancer H3K4me3 devoided regions are highlighted in gold, while well located nucleosomes are shaded in purple. b) Gene body extension of H3K4me3 in TSGs. c) Potential higher methylation levels for aberrantly expressed OGs.

As a summary, in this section we showed the suitability of CNNs to analyze the properties of H3K4me3 enrichment profiles. Yet, it was not clear if the use of a deep neural network was in fact necessary, or if it would constitute an improvement from more conventional methods. Whether a simpler approach would be satisfying enough remained to be checked. To do so, a Principal Component Analysis and Clustering of the signals were conducted.

4.2 Principal component analysis and clustering

Principal Component Analysis (PCA) and subsequent clustering of the cropped H3K4me3 signals around the promoter were performed for all cancer tissues and their respective matching healthy samples. We included in the study both Training-Validation (subset A) and the Test (subset B) genes. For simplicity and without loss of generality, we will discuss and illustrate the results using Lung I and Liver I as references.

This early step in the data analysis pipeline was conducted with the idea to bring to light possible fundamental differences between sequences corresponding to different categories, for example cancer driver genes vs. housekeeping genes, or healthy vs. tumor samples. PCA analysis would in principle provide some insights regarding the nature of the signals and allow us to assess the difficulty of a classification task using H3K4me3 ChIP-seq data.

Pursuing to explore how much information is encoded in the main principal components, scree plots with the eigenvalues of the respective covariance matrices and the cumulated variance ratio were plotted in Figure 8. The eigenvalues of the covariance matrix quantify the relevance of each new direction or principal component, while the cumulated variance describes the percentage of information covered by the main components.

Scree plots for both cancer and healthy samples are shown at the leftmost part of Figure 8. These curves present the main elbow, the point where the slope of the curve changes the most, close to the third component. In a normal picture that would imply that the main three components should be enough to describe our sequences.

However, the second plot in Figure 8 shows that this is not the case. The first ten components together did not reach to explain a standard minimum of 80% of the total variance, meaning that not even the main principal components can synthesize our data properly. It is important to note that, independently of the amount of variance encoded in the principal components, one cannot infer anything about the separability of the signals.

The signals were then represented in the three dimensional space defined by the first three principal components. As a representative example, the results obtained from the Training and Validation sequences for Liver I are presented in Figure 9.

Figure 9a) shows that the studied cancer samples are characterized by a lower variance than the ones obtained from healthy tissue. This phenomenon, observed even after sequence normalization, agrees with the results presented in Figure 8. This explains the lower values for the eigenvalues in cancer samples, and why the curves for these samples generally lie below their matching healthy counterparts. As a consequence, sequences obtained from the cancer tissue are clustered in a narrower region within the PCA space. This region, nevertheless, clearly overlaps with the area occupied by the sequences from the healthy sample.

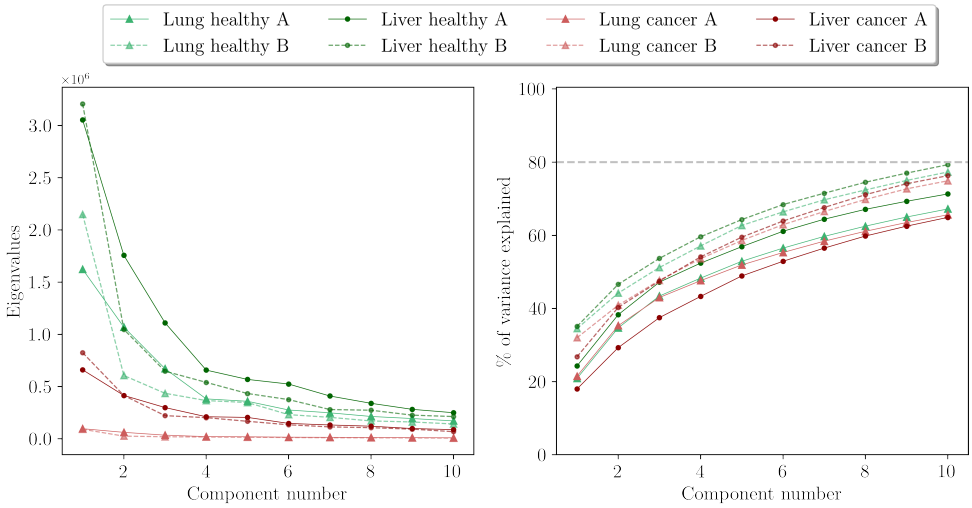


Figure 8: **Principal Component Analysis**. Scree plots for healthy and cancer samples, showing the eigenvalues of the covariance matrix. Cumulative variance vs. component shows the amount of information encoded until a certain PC.

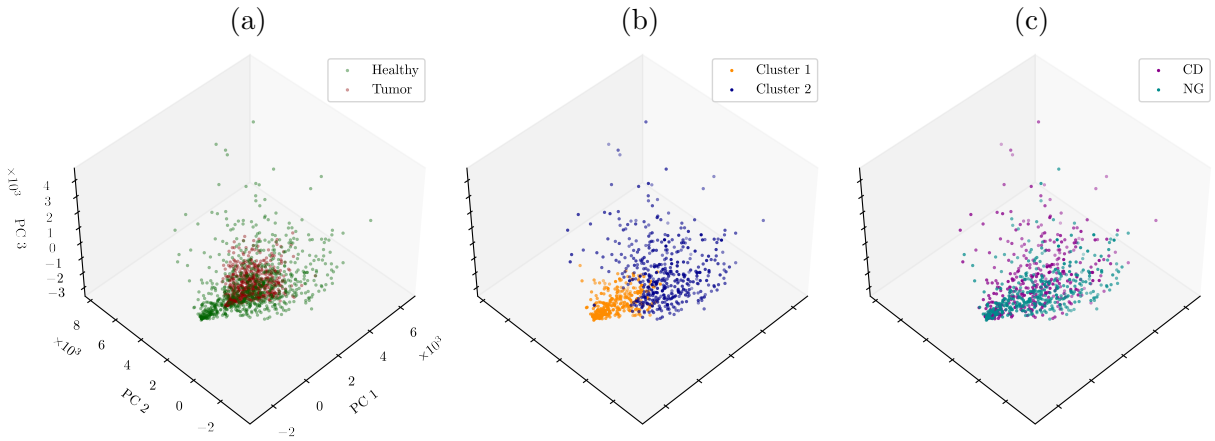


Figure 9: **Sequence representation in PC space** a) Cancer vs. healthy samples. Cancer samples cluster in a narrow region within the volume occupied by healthy samples. b) Clustering algorithm data classification for sequences from the healthy sample. The algorithm splits the data in two regions geometrically. c) Cancer drivers vs. neutral genes. CDs and NGs appear mixed in PC space.

Figure 9b) illustrates that, if there are any intrinsic differences between the sequences corresponding to different target categories, they do not appear when performing a standard PCA. To ease visualization, only sequences from the healthy sample are shown in Figures 9b) and 9c). Several clustering algorithms were tested, including K-means, hierarchical clustering and agglomerative clustering, which is the one shown in Figure 9b). None of these methods were able to find meaningful boundaries in the low-dimensional representation of the data, but rather proposed geometrically based classes that had no biological relevance.

Figure 9c) confirms that genes cannot be classified by clustering their H3K4me3 profiles. In this figure, which is the most important for our purposes, the data points for each sequence are filled with a color representing their ground truth labels, i.e. CD or NG. The main conclusion one can extract from this plot is that cancer drivers and normal genes seem to be mixed in the PCA space, with no apparent geometrical way to classify them. Nonetheless, even though there are many CDs in the bulk of NGs, a significant number of them spread in the directions of PC2 and PC3 for this particular subset of genes.

From these results and the information gathered from a visual inspection of the signals in the previous section, one can already obtain the following meaningful conclusions.

First, a linear transformation of the data is not enough to capture properly the nature of our dataset. The performed PCA analysis strongly suggests the need to go deeper in the level of complexity of our approach. Second, the H3K4me3 enrichment profiles for oncogenes and tumor suppressors will be genuinely similar to those characterizing housekeeping genes.

This means that the information concerning the category of the gene that could be encoded in the sequences, if any, could involve "how" the features are used or "where" in the signals do they appear. The possibility that the differences between CD and NG profiles involve the use of completely different patterns seems unlikely.

4.3 Introducing artificial neural networks for the analysis of ChIP-seq signals for histone PTM enrichment.

Our preliminary results suggested the potential of CNNs to extract information from H3K4me3 enrichment profiles[8], proved the suitability of CNNs to be used in the cancer gene classification task, and illustrated the necessity to go deeper in the complexity level of our approach to the problem.

The above considerations inspired the creation of a customized deep convolutional neural network, OriGENE, presented in the following section.

4.3.1 Architecture

At an early stage, Recurrent Neural Networks (RNNs) were considered given their widespread use in sequence analysis and motivated by the fact that genes vary in length. This research path was abandoned for several reasons: 1) *Online learning*, i.e. updating the model's weights sample to sample, would be the only way to exploit the potential of RNNs to handle inputs with varying sizes. Zero padding of the sequences would otherwise be required in order to train in batches. 2) There is only a significant enrichment of H3K4me3 at the beginning of the genes. Cropping the signals to a standard size seemed thus to be justified. 3) The patterns we are analyzing are static and 4) a final fully connected classifier would already take into account their spatial distribution and order of appearance.

Given these facts, using a convolutional-based method was the most logical way to proceed.

The goal was to develop a CNN tailored to the needs of the project, especially given the use of 1D sequences. The ideal architecture should be complex enough to have the potential to identify subtle trends and non-trivial details in the data. It should also have the capacity to integrate them to give accurate predictions. At the same time, the number of trainable parameters of such architecture should remain low enough not to over-fit the model to the reduced dataset.

These basic principles guided our exploration of the hyperparameter space, i.e. the possible combinations of layers, filters and regularization elements. The best trade-off was achieved with the architecture we propose in the next section.

[8] As a reminder, these profiles illustrate the significance of a certain histone post-translational modification (PTM) like H3K4me3 throughout a genomic region. The profiles were obtained using Chromatin-ImmunoPrecipitation and Sequencing (ChIP-Seq).

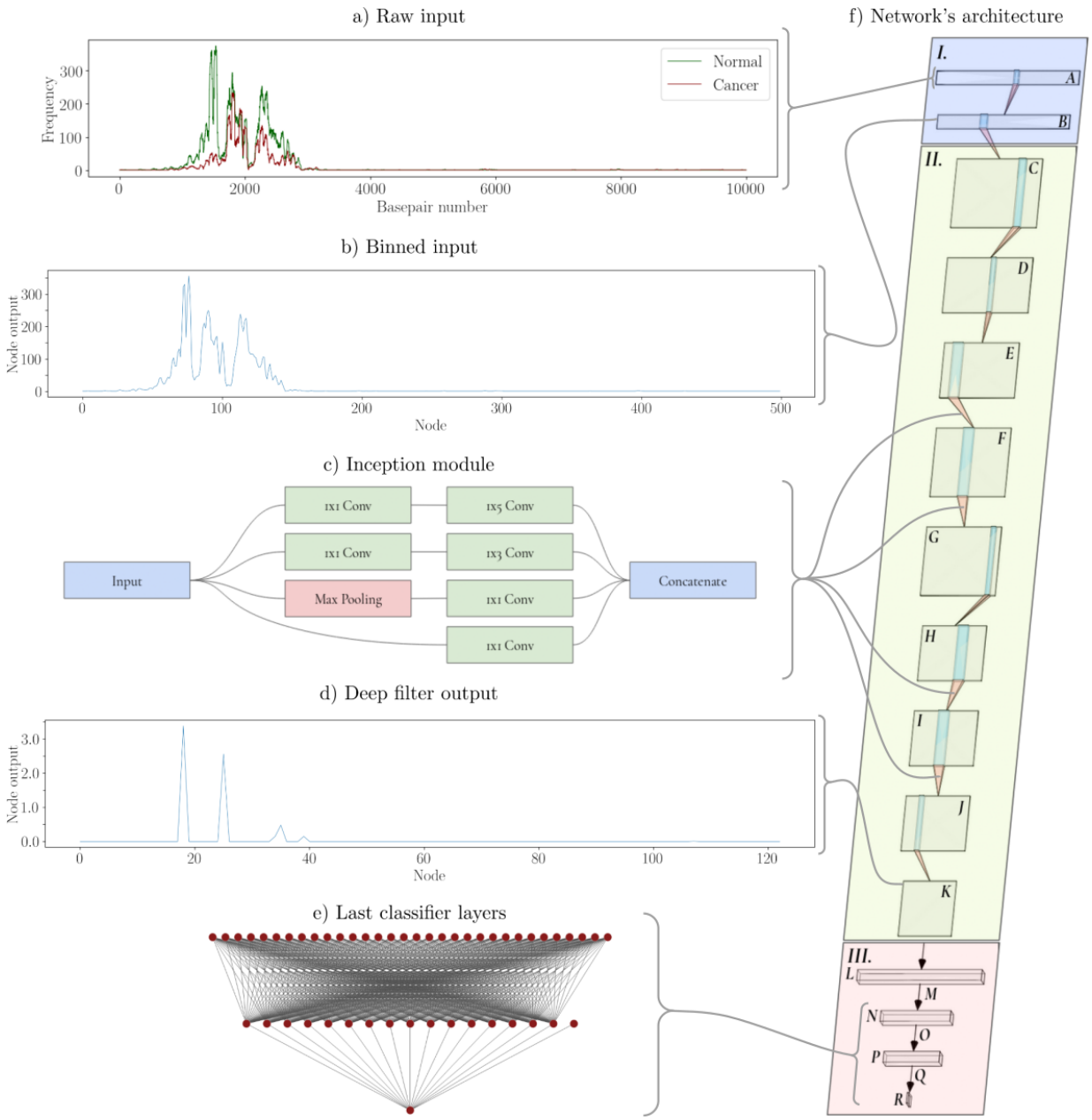


Figure 10: **OriGENE illustrated**. a) Input signals before and b) after being binned for the healthy track, with a 20 bp resolution. c) Inception module structure, with convolutions of kernel one aimed to lower the number of trainable parameters, followed by different size kernel convolutions that process information and study signal features at different spatial scales. d) Output of one of the 16 final filters, which will be flattened and fed into e) the final fully connected layers, which will perform the classification. f) Shows the global architecture of our model, with the different blocks and layers specified.

Table I. The architecture of OriGENE				
	Layer	Type	Output shape	Parameters
	I. Input			
A	**Original signal**	*Input layer*	(None, 10000, 2)	0
B	**Binned signal**	*AveragePooling1D*	(None, 500, 2)	0
	II. Feature extractor			
C	**Basic Convolutional (I)**	*Conv1D*	(None, 249, 32)	224
D	**Basic Convolutional (II)**	*Conv1D*	(None, 247, 16)	1552
E	**Size reduction (I)**	*AveragePooling1D*	(None, 123, 16)	0
F	**Inception A (I)**	*Inception*	(None, 123, 32)	1072
G	**Inception A (II)**	*Inception*	(None, 123, 32)	1584
H	**Size reduction (II)**	*MaxPooling1D*	(None, 61, 32)	0
I	**Inception B (I)**	*Inception*	(None, 61, 16)	664
J	**Inception B (II)**	*Inception*	(None, 61, 16)	408
K	**Size reduction (III)**	*MaxPooling1D*	(None, 30, 16)	0
	III. Classifier			
L	**Flattened array**	*Flatten*	(None, 480)	0
M	**Dropout (I)**	*Dropout*	(None, 480)	0
N	**Fully Connected (I)**	*Dense*	(None, 32)	15392
O	**Dropout (II)**	*Dropout*	(None, 32)	0
P	**Fully Connected (II)**	*Dense*	(None, 16)	528
Q	**Dropout (III)**	*Dropout*	(None, 16)	0
R	**Last Fully Connected**	*Dense*	(None, 1)	17

Table 1: **The architecture of OriGENE**. Three distinct blocks characterize the network. I. Input, II. Feature extractor and III. Classifier. The output shape of each layer (a,b,c) describes a) the number of paired samples used, which depends on the stage of the procedure, b) the sequence length and c) the specific number of outputted channels or filters. A significant dimensionality reduction takes place before the third block, while most of the trainable parameters come from the first fully connected layer.

4.3.2 The model: OriGENE

The architecture of OriGENE[9], illustrated in Figure 10 and described in Table 1, consists of the following three blocks:

I. Inputs.- The use of paired tracks for healthy-tumor matching samples as aligned and parallel inputs seemed promising, since the objective is to find genes linked with cancer. This method was favoured instead of using a higher number of epigenetic markers for unmatched samples.

Our second conclusion was that it was better to study in detail the region surrounding the TSS, where transcription of the gene starts, than trying to analyze a broader region. As shown in previous sections, there is only a significant enrichment at the promoter of active genes [27], although H3K4me3 can also appear in DNA-repair regions. However, the signal in the gene body is considerably milder than around the promoter and we lack the genetic context of the said region.

The above points left us with two aligned sequences of 10.000 basepair enrichment values, starting around 2000 bp before the TSS of each gene.

The desired resolution of the input sequences constituted a first degree of freedom for our network. In spite of counting with enrichment values at a basepair level, the effective resolution of the raw signals was indeed lower. The smallest relevant details starting to emerge at a scale of 20-50 bp.

The original inputs were hence binned in 20 bp averaged blocks before extracting any features from them. This step contributed to a remarkable size reduction of the sequences, as seen in the first block of Table 1.

II. Feature extractor.- The general design of the feature extractor introduced a level of arbitrariness. Starting with a high number of filters and reducing it as one goes deeper into the network, in an inverted pyramid-like shape, proved to be useful in this particular project for the following reasons.

First, the network needs to collect as many low-level features from the original signals as possible in order to combine them in later steps. The algorithm starts from scratch to analyze and process the almost raw enrichment profiles. Therefore, the first filters will play an important role in identifying basic properties. Given that we were comparing just two 1D sequences, there was no need for the moment to count with more filters at the deepest and most abstract layers.

Second, having fewer filters to flatten at the deepest layers of the feature extractor shortened to a great degree the vector that constituted the input to the first fully connected layer.

[9]Given that the network is based on the INCEPTION module, named after a movie title translated as Origen in Catalan and Spanish, I decided to call the model OriGENE.

The inception module was introduced as the core convolutional element in this second block of the network. This was motivated by the necessity to analyze the structure of the signals at varying spatial frequencies. The module, presented for the first time in [31], is constituted by a small set of pooling and convolutional filters of varying kernel size that enable it to find and combine the information at different spatial scales and resolutions. The inception module is further detailed in Appendix B.1.

As a common practice, several pooling layers were added between the different inception modules. These were aimed to further reduce the size of the data before proceeding to integrate all the extracted features in the final stage.

III. The classifier.- The last filter outputs are flattened and a set of fully connected layers integrates all the extracted features. The classifier gives a unique output for the final node: the probability of the input sequences to be attributed to a cancer driver gene.

These last fully connected layers were in fact the main contributors to the number of trainable parameters, as shown in Table 1. They were thus proportionally regularized using dropout (See Appendix B.1).

Unlike the original INCEPTION network presented in [31], where only one last dense layer was used in the main branch, the use of more than one fully connected layer before reaching the final node was justified in our problem. The spatial and sequential distribution of the found patterns could play a role in H3K4me3 signals, and hence it had to be taken into account by allowing connections between different regions of the final vectors.

4.3.3 Benchmarking OriGENE

Our findings can be partially benchmarked using the data and results presented by Jie Lyu et al. in Table 1 of [11].

Their study, in which the performances of different models were tested when given OGs and TSGs from the COSMIC's Cancer Gene Census (CGC) v.87 [21], showed that most of the available algorithms by 2020 ([1]-[10]) were proficient at identifying genes that were labelled as "neutral", since their performances were characterized by truly remarkable specificities. Nevertheless, once a false-positive discovery rate below 1% was imposed, even DORGE [11] had a limited ability to filter out the actual oncogenes and tumor suppressors, with a combined sensitivity that without a proper context would seem to be rather low[10].

Their results give us an idea of the difficulty level of the problem we are trying to solve, and are the ones that will be used as a reference to assess the potential and limitations of the method we propose.

[10]DORGE achieved the best sensitivity of all the presented algorithms, 0.611, when predicting CGC genes at an imposed 1% false-positive rate.

4.4 Binary classification problem: Cancer Driver vs. Neutral Gene

The project was originally conceived as a three-class classification problem, with the categories OG, TSG and NG. Such a procedure would allow the network to reuse and combine features from different classes simultaneously, if necessary, making the learning process more efficient.

The labels TSG and OG are however non-exclusive attributes, which can in addition be tissue- or sample-dependent. These categories, when referring to the functional role of such genes, should therefore not be an intrinsical property of a region in the genome regardless of the context. As an example, unmutated OGs operating normally are called proto-oncogenes and are necessary for the well-functioning of the cell.

This subtlety implies that 1) the labelling scheme already introduces errors into the system in a pan-cancer study, since the same gene can be a TSG in a cell line but an OG in another, and 2) increases the need of a significant amount of samples for each class for an algorithm to be reliable.

A tissue-specific approach could in principle solve point 1), but the data in our first reference and carefully curated dataset, the CGC [21], was scarce enough not to justify a three class classification algorithm in a tissue-specific context at that stage of the project. Besides, the realization that different datasets could not be integrated in our training process limited even further the data we could work with.

These constraints led us to adopt a less restrictive approach, and ask whether the genes with the given H3K4me3 profiles could be causally related to cancer genesis or not. OGs and TSGs were put under the umbrella category of cancer drivers (CDs), and the number of NGs was doubled in order to keep the class balance, since the known housekeeping genes outnumber the known CDs.

4.4.1 Model development: Training and Validation

The first 190 OGs, 190 TSGs and 380 NGs stated in the supplementary data table S2 of [11] constituted our Training-Validation set. In order to make the most of the available data, the corresponding 760 pairs of samples (H3K4me3 aligned tracks for healthy and matching cancer tissues) for these genes were distributed randomly in 10 stratified splits, aiming to keep the class balance. We then proceeded to perform K-Fold validation of several candidate network models[11].

[11] In every fold, 76 pairs of sequences were hidden to the network and saved for validation, while the remaining 684 pairs were used to train it. The networks learned the best when training in batches of 98 samples, which smoothed the contributions of noisy sequences to the weight's updates, throughout 100 epochs. Weights were updated 7 times per epoch, corresponding to 6 minibatches of 98 paired sequences, and a last one that contained 96. Within the chosen number of epochs one could not see signs of overfitting to the training set.

Designing OriGENE. Rather early in the development stage we could notice that the limited number of known genes linked to cancer narrowed the range of reasonable sizes for our network. Hence, we dismissed the possibility of basing our model on complete and sophisticated deep networks such as the ResNET.

The introduction of the inception module helped to ease and smooth the training process. This can be seen in the stable loss curves and slow but rather smooth learning shown in Figure 11a). After 100 epochs, the loss curves started to flatten, although not completely. The accuracy for both validation and training sets, though, stagnated after this point.

The use of inception also reduced significantly the amount of trainable parameters. This was a consequence of the module using size one convolutions. After tuning model hyperparameters such as the number and type of layers, regularization elements such as L2 or dropout, etc. we obtained our final model (See Appendix B.1 for more information regarding the inception module and hyperparameter tuning).

Validation remarks. The predictions of OriGENE for the different validation splits were then joined. The corresponding Receiver Operating Characteristic (ROC) curves with their Area Under the Curve (AUC) values were calculated and plotted in Figure 11b). This procedure was performed for each patient individually. The robustness and reliability of the models trained on different individuals were estimated by AUC 90% confidence intervals, which were assessed by bootstrapping 1000 times from the predicted class probabilities and analyzing the corresponding AUC value histograms, as seen in Figure 11c).

A thorough analysis of the training results and validation performances, which in principle should give a hint on what to expect when testing new data, already unravelled several facts.

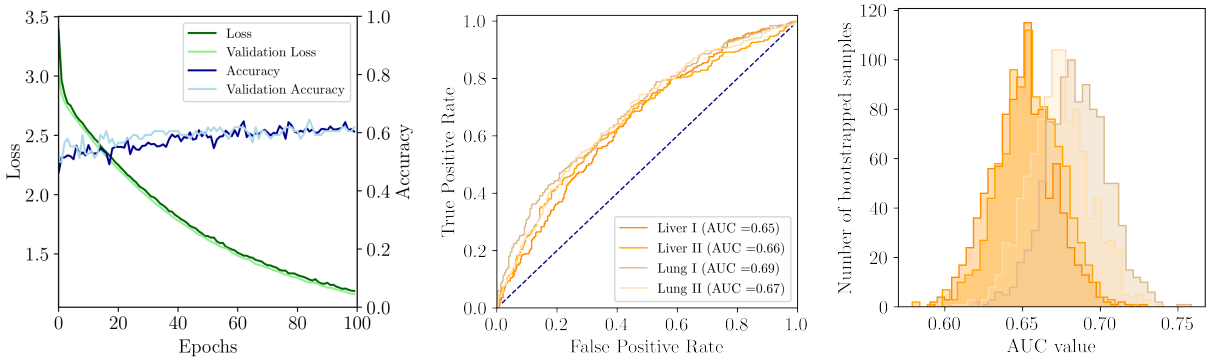
First of all, we realized that no matter how complex the network was, a 100% classification accuracy for the training data was never achieved without clearly overfitting the data. Secondly, the performance of the model in different folds was highly sensitive to the splitting scheme, an effect that was smoothed when combining all validation predictions.

These phenomena, which would also impact the following steps, are unavoidably enrooted in the nature of our approach and deserve to be studied in some detail.

Upper performance bounds and fold dependence. The labelling scheme and the input sequences were intrinsically noisy, as it was seen in Section 4.4 and discussed in Appendix B.2.1 respectively. Such noise sources would constitute a first constraint on the top performance our models could achieve.

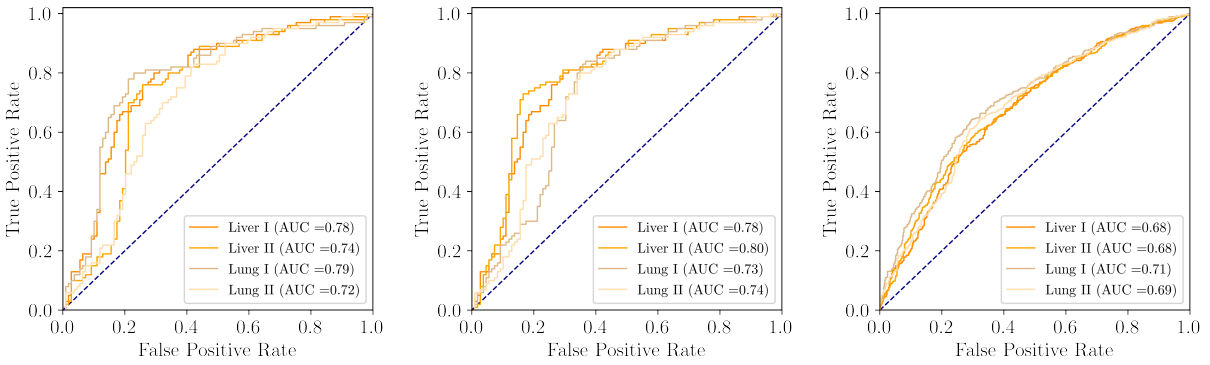
In addition, our network had the potential to pinpoint CDs from only a subset of types[12]. Our approach would only enable the network to find the CDs that can leave a distinct fingerprint in the H3K4 trimethylation profiles around the promoter and beginning of the gene body. According to [11], [12], [17], [18] and [32] though, using H3K4me3 as the

[12]In [12], CDs were classified according to their nature in: interactome-driven, mutation-driven, DNA-methylation driven, expression-driven and copy number aberration-driven.



(a) (b) (c)

Figure 11: **Validation summary plots.** a) Learning curves and loss evolution for one validation fold example. Slow but steady learning curves were achieved during training, with validation and training loss curves following similar trends due to the introduction of the inception modules. After 100 epochs the accuracy stagnated. b) ROC curves with the combined validation predictions for all the available tissue samples. Similar ROC curves for OriGENE predictions are obtained when the network is trained on different datasets. c) AUC histograms for the validation bootstrapped predictions. Rather narrow distributions led to tight confidence bounds for single model predictions.



(a) (b) (c)

Figure 12: **Test ROC curves**. a) Original test ROC curves when predicting and training with data from the same patient. b) ROC curves for OriGENE trained on Liver I predicting the test set for other patients and tissues. c) ROC curves for final assessment loop predictions, using all data split randomly.

unique marker could introduce in our current procedure a bias towards TSG detection. Despite the marker having a certain prediction power for OGs, they are known to be better characterized by other elements such as promoter and gene body DNA-methylation aberrations, H3K27me3 silencing domains or DNA mutations.

An uneven distribution of the targetable CDs could hence be at the core of the observed significant fold-dependent performance fluctuations. The observed split-dependent behaviors were further enhanced by the limited number of sequences. This last contribution can be compensated in several ways, e.g. combining the predictions for different folds as it was done here.

The potential of OriGENE. In spite of the already known limitations of the model, we started to observe the main conclusion from our work: OriGENE could effectively learn from all datasets independently.

OriGENE showed way above random ROC curves for all tissues, presented in Figure 11b), and AUCs ranging from 65 to 68 with narrow confidence intervals, presented in Figure 11c). Not only the model seemed to retrieve useful information from the sequences, but according to the validation performances the model could in principle achieve significant sensitivities. The model's specificity was finally the best performance metric in this early stage of the pipeline, pointing towards a better ability to distinguish NGs than CDs that is shared with all the algorithms presented in [11]. These results are summarized in Table 2, and will be further confirmed and elaborated in the following section.

4.4.2 Model evaluation: Hold-out testing

Once the final OriGENE model was fully designed and settled, the model's weights were trained using all the training-validation sequences. Its ability to generalize to unseen data was then assessed using a completely independent test set constituted by 109 NGs and 100 CDs.

The results obtained in this case are shown in Figure 12a). The presented ROC curves are characterized by AUCs between 70 and 80%, and rather narrow confidence bounds. High specificity values, showing the notable contribution of the correct NG classification to the model's performance, can be highlighted from the second block in Table 2. This trend agrees with what was observed in the model's validation stage presented in Section 4.4.1.

The fact that the performance in the test set was indeed better than that of the validation stage further corroborated the dependence of the model performances in the way the data is split. Our arbitrary and uninformed original choice of test sequences had been, at least, unfortunate.

On the other hand, these results constituted a strong proof that the model had the potential to predict previously unseen data. Unlike the validation sequences, which had been used to fine tune the model's architecture and hyperparameters, the new test set was constituted by genuinely independent genes that had not played any role in the model selection process.

The model's capability to extract utile and generalizable information from the almost raw H3K4me3 profiles could not be doubted anymore, and further justified our project's previously unexplored approach.

Ensemble averaging. Every time a model is trained, even with the same data, it ends up with completely different sets of weights that can perform the classification task in a similar fashion. We noticed that in some runs the correct classification of negative samples slightly dominated the training, while sometimes the opposite picture arised.

In order to counteract the effects of different weight initialization maps, ensemble averaging was performed (See Appendix B.1). The model was hence trained 10 times for each patient, and the classification outputs of single sets of weights were averaged aiming to obtain a more accurate prediction for each test sequence pair. Ensemble averages are known to add an extra level of regularization, and ensure that the final performance will be as good or even better than the averaged predictions of single models.

The results obtained from this sanity check, which induced little to no changes in the original performance metrics' values, are the ones shown in Table 2 and Figure 12a). This step made us discard weight initialization as the source of the test performance improvement.

Cross-patient and cross-tissue prediction. An additional feature of OriGENE that could also be tested was the ability of a model trained on a specific dataset to predict samples of a different tissue or patient origin.

Models trained on Liver I and Lung II displayed an outstanding capacity to cross-predict genes. The ROCs for Liver I shown in Figure 12b) corroborate it. This result provides additional support to our first conclusion, namely that the incredible variability among the trimethylation patterns of different genes is undercovered by a more than substantial similarity of same-gene profiles, even in different individuals.

The apparent limitation of the approach we followed turned out to be one of the properties that could make it appealing for its generalizability.

Table II. Summary of the main performance metrics							
Stage	Tissue	Acc	Sp	Sn	Prec	AUC	90% CI
Validation	*Liver I*	0.607	0.655	0.558	0.618	0.650	[0.616 – 0.684]
	Liver II	0.618	0.695	0.542	0.639	0.655	[0.622 - 0.687]
	Lung I	0.618	0.663	0.574	0.630	0.685	[0.653 - 0.716]
	Lung II	0.626	0.637	0.616	0.629	0.673	[0.641 - 0.708]
Test	*Liver I*	0.675	0.872	0.460	0.767	0.782	[0.723 - 0.834]
	Liver II	0.536	0.844	0.200	0.541	0.742	[0.682 - 0.798]
	Lung I	0.608	0.890	0.300	0.714	0.791	[0.735 – 0.843]
	Lung II	0.684	0.734	0.630	0.685	0.717	[0.658 – 0.775]
Looped test	*Liver I*	0.633	0.632	0.633	0.628	0.676	[0.647 – 0.704]
	Liver II	0.636	0.622	0.650	0.628	0.683	[0.655 - 0.712]
	Lung I	0.670	0.693	0.646	0.674	0.710	[0.683 - 0.736]
	Lung II	0.657	0.654	0.660	0.652	0.687	[0.659 - 0.714]
Cross-tissue prediction	*Liver I*	0.675	0.872	0.460	0.767	0.782	[0.723 - 0.834]
	Liver II	0.603	0.899	0.280	0.718	0.797	[0.742 - 0.849]
	Lung I	0.656	0.734	0.570	0.663	0.726	[0.668 - 0.784]
	Lung II	0.675	0.771	0.570	0.695	0.736	[0.676 - 0.793]

Table 2: **Performance metrics for the different stages in the procedure**. Values were rounded to the third decimal cypher.

Table III. Liver specific performance metrics								
Stage	Tissue	Thr	Acc	Sp	Sn	Prec	AUC	90% CI
Tissue specific	*Liver I*	0.61	0.616	0.877	0.356	0.743	0.639	[0.605 - 0.674]
		0.50	0.618	0.655	0.580	0.627		

Table 3: **Liver specific performance metrics.**

4.4.3 K-fold cross testing

Given the unexpected better results for the test set than the ones obtained in the validation stage, a further evaluation and compensation of the bias introduced by the original arbitrary splitting strategy proved to be crucial at this point of the project.

Estimating the model's prediction power. An additional K-folded loop with all the available data was conducted as a final and more reliable estimate of the actual prediction power of the model. All sequences were now randomly split in five stratified blocks, of which four where used to train and one to test. All folds' test results were then gathered for the final evaluation as done in Section 4.4.1.

It should be noted that this procedure was not fully devoided of bias. Data used to train and validate before, which played a role in the model selection stage, is now being tested. Therefore, there could be an information leakage. Besides, the training sets now contained 775 sequences in order to split the dataset in rather uniform batches, meaning that the network had slightly more available data[13].

Albeit the aformentioned problems, the proposed protocol would ensure that each training set had enough samples to learn from, while counting with a large effective test pool that would counteract the effects of limited dataset sizes and partition-induced imbalances.

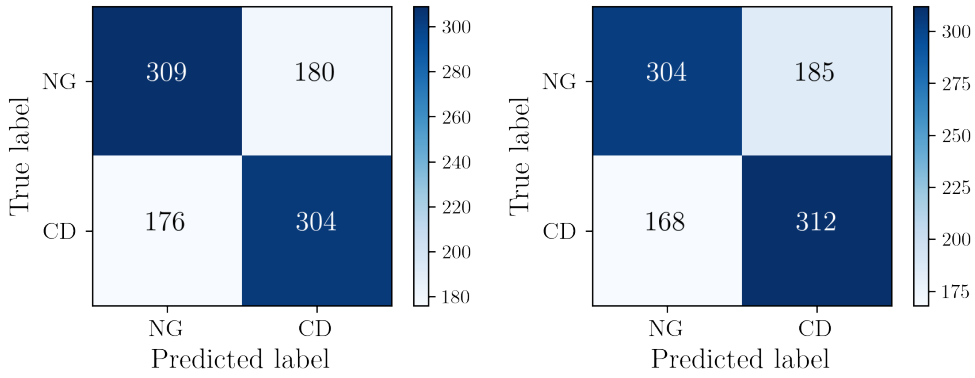
Final evaluation in a pan-cancer scenario. Our last results are presented in Figure 12c), the confusion matrices presented in Figure 13, and the performances in the third block of Table 2. AUCs close to 70% with narrow confidence intervals characterized once more the ROCs obtained. Remarkably, this final evaluation of the model showed a slight drop in specificities, which made way for sensitivities above 60% while keeping the precision high. Regarding this second critical performance metric, the model's precision, a minimum of 6 out of 10 cancer driver predictions would be correct independently of the patient and tissue used to train.

It is of key importance to point out that sensitivity and precision are the metrics one wants to enhance. The main goal of our work is to discover new cancer drivers, while the confirmation of already known neutral genes is desirable but remains in the background.

The final precision and sensitivity combined values for OriGENE prove that, when using the same CGC v.87 CDs and a subset of the further curated NGs from [5] used in the evaluations presented in [11], the proposed model outperformed other well-established genome-based algorithms in the binary classification framework. According to [11], only DORGE displayed a better performance when filtering out actual cancer drivers from the chosen database than the model proposed in this thesis.

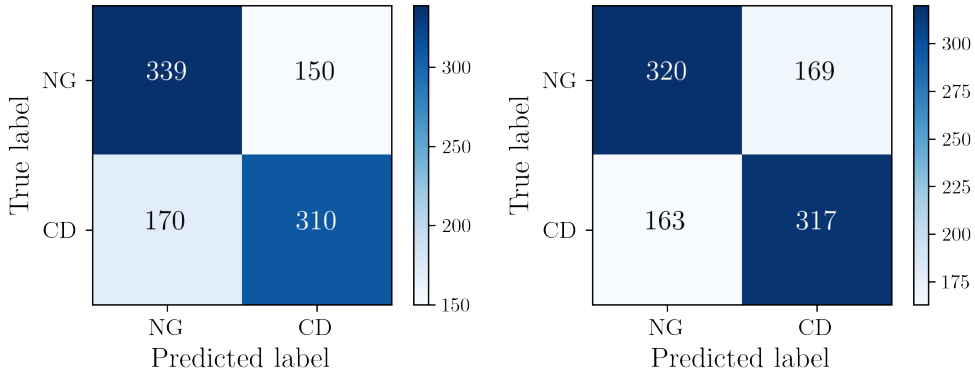
The precision-recall curves and the related average precision values, which correspond to the Area Under the Precision-Recall Curve (AUPRC), are shown in Figure 14. These curves were plotted to ease the comparison with other models, in particular DORGE.

[13] 15 extra sequences, 2% more than the original dataset



(a) Liver I

(b) Liver II



(c) Lung I

(d) Lung II

Figure 13: **Final test confusion matrices setting the cutting threshold at 0.5**. K-fold loop on whole dataset split in fully randomized and stratified blocks. Performance metrics are shown in Table 2

DORGE-TSG and DORGE-OG, when using only epigenetic predefined features and all NGs available, achieved AUPRCs of .6 and .295 respectively. OriGENE reached combined AUPRCs of .65 or greater in all samples with a NG-to-CD ratio of $\sim$1:1.

Although these numbers cannot be directly compared due to the impact of varying class ratios in the precision values, they would support our working hypothesis. The proposed unconstrained feature extraction process for H3K4me3 enrichment tracks could in fact constitute an improvement when comparing with the manual curation of features. These results would motivate further research in the same line.

4.5 Tissue-specific prediction

The project was lastly redirected towards the discovery of tissue-specific cancer-driver genes.

The ability to work in a tissue-specific picture is an extremely desirable feature for a cancer gene classifier that OriGENE could incorporate naturally. Single cancer-healthy matching pairs of samples were already available from [17] and we counted with an architecture that could analyze them. The only element to be updated was the specific pool of genes, or more precisely the sequences and their corresponding labels, that could be used to train the network.

For this purpose, cancer drivers that were known to play a role in liver cancer, angiosarcoma, carcinoma and lymphoma were obtained from Cancermine [22] and extracted from Liver I samples. The collated table in Cancermine contains a list of cancer driver genes that have been cited in at least one publication. The 374 new liver-specific CDs were compared with 374 NGs following the same procedure as in Section 4.4.3. Data was now split in 9 folds and trained in batches of 83 sequences.

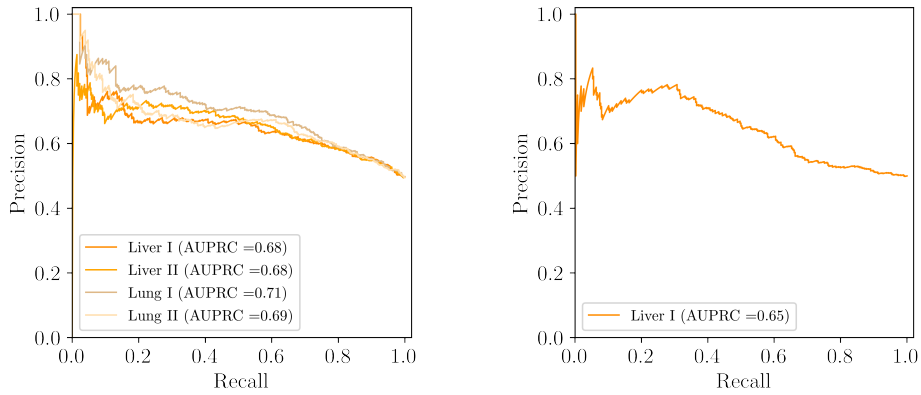
We noticed a slight increase in the number of noisy and low enrichment level sequences with a certain structure for the new pool of positive genes. A fact that would explain the nature of some of the low enrichment sequences would be a higher number of reads being mapped to secondary alignment hg38 domains for the new set of CDs[14]. Whether the new CDs are expected to be active or not could also play a role in the H3K4me3 profiles, since this PTM is known to be a mark of active genes.

Our preliminary results are presented in Figures 14b) and 15 and described in Table 3. These results support the robustness of OriGENE with a performance that was not compromised when changing to a tissue-specific context.

Specificity had still the best values, while sensitivity and precision were somewhat better when compared with the validation stage of the pan-cancer assessment. In addition, the false-positive discovery rate could be lowered to 12.3% while keeping OriGENE's sensitivity in a tissue-specific scenario above that of all the only-genomic models presented in [11] when using CGC genes in a pan-cancer scheme.

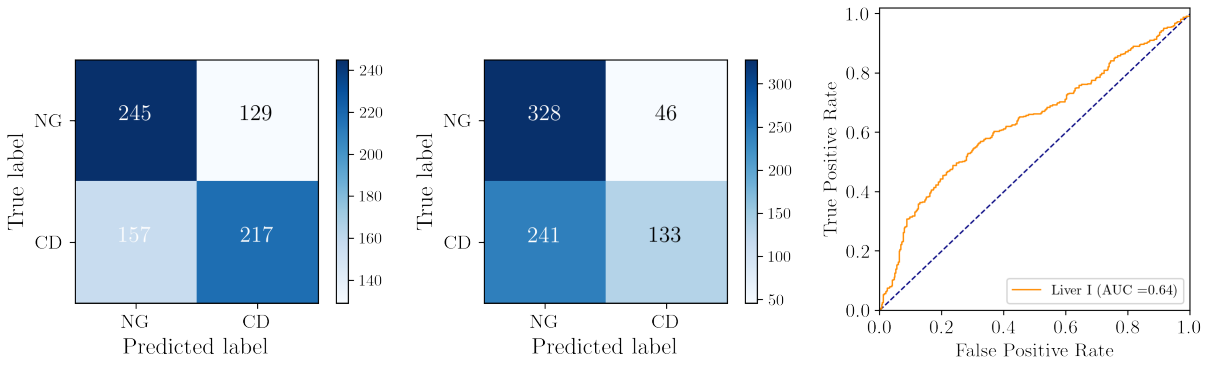
While the ability to discard neutral samples was kept to the same level, the network could also distinguish CDs that showed the expected average levels and profiles for both tracks. CDs where the cancer track clearly dominated in high enrichment regimes were also pinpointed by the network. More interestingly, it classified correctly positive sequences where one could observe mild but sustained enrichment levels and small blobs for the tumor track in the first part of the gene body.

[14]Gene coordinates were queried using The Genome Browser [33] for each gene. The domains related to single genes had multiple entries with non-unique coordinate limits, including secondary alignments for hg38. Some reads could hence be lost when mapping them to different sections of the reference genome.



(a) Testing loop. (b) Liver-specific.

Figure 14: **Precision-Recall curves and AUPRCs.** Final test loop and liver-specific Precision-Recall curves with the corresponding areas under the curve, which can be understood as the average precision.



(a) 0.5 threshold (b) 0.61 threshold (c) ROC

Figure 15: **Liver specific performances**. a), b): Confusion matrices setting the decision thresholds at 0.5 and 0.61. c) ROC curve for the prediction of liver-specific genes.

If these features were not artifacts, e.g. background noise induced by the lack of a non-specific input, they would motivate the need to efficiently extend the network to the body of the genes. In this region the known gain- and loss-of-function (GoF, LoF) mutations for OGs and TSGs respectively could modify the histone PTM environment locally.

Further evaluation of OriGENE in this last scenario is required, but the potential of the model to handle distinctive properties of a single tissue, in this case liver, and to tailor the weights of the already designed network to specific patients and cancer types, holds promise and encourages us to continue with this research path.

5 Conclusion

The enrichment profiles of the epigenetic marker H3K4me3 around the promoter of active genes encode both functional and structural information that can be useful for the characterization of cancer driver genes. These profiles constitute a signature of each gene. Therefore, sequences corresponding to a certain gene, even if they belong to different individuals or they have a different health status, cannot be considered completely independent. Inter-gene variability is more significant than the intra-gene feature differences, and this must be considered when designing a cross-validated project.

Principal component analysis (PCA) of the samples could not represent the data efficiently, given the need of many principal components to capture a reasonable amount of variance. Moreover, several clustering algorithms including K-means and hierarchical clustering failed to find meaningful boundaries in the low-dimensional representation of the data, highlighting the need to go further in the level of complexity and abstraction of our study.

The aforementioned preliminary results motivated the development of OriGENE, a deep convolutional neural network based on the inception module aimed to analyze, compare and bring out features from matching healthy and cancer samples. OriGENE introduces a novel unconstrained, multi-scale and spatially aligned feature extraction process, which allows it to process information down to a 20 bp resolution.

Several factors imposed an upper bound on OriGENE's performance. First, the network had to be trained on a limited number of highly curated genes using single tissues. Second, the labelling scheme and the sequences themselves were intrinsically noisy. Last, H3K4me3 had a lower ability to predict oncogenes than tumor suppressors, which led to a pronounced fold dependence. This behaviour was compensated and smoothed by adding different stratified and randomized fold predictions together.

These limitations provided the perfect environment to explore new properties of our model, namely its potential to be trained and predict CDs using different tissues, but also the capacity to operate in a cancer-specific context, with both attempts being fruitful. Even though OriGENE displayed on average better specificities, it stands out by a noteworthy sensitivity to precision relation, i.e. a significant number of actual CDs are detected (60%+ in the final pan-cancer evaluation of the model on CGC [21] genes), combined with a still higher than 60% chance of the predicted CDs to be correct. OriGENE would also show promise in the tissue-specific and patient-specific front, since the presented model can already be trained but also predict single-tissue data with the same reliability as in the pan-cancer picture.

All in all, the obtained results consolidate OriGENE as a model with performance strengths that are complementary and comparable to other already well established algorithms such as OncodriveFM [1], MuSIC [2], MutSigCV [3], OncodriveCLUST [4], TUSON [5], ActiveDriver [6], 20/20+ [7], OncodriveFML [8], MutPanning [9] and GUST [10], the performances of which were presented in DORGE [11].

We hope that this project, born as a proof of concept, will serve as a precedent for future studies merging deep CNNs and cancer gene prediction.

Outlook

This section's purpose is to suggest new ways to expand the scope of the project, and to discuss the possible impact of the proposed method.

It would be of interest to extend the studied regions further into the gene body, where the most studied somatic mutations take place. The addition of mutation sites and types as complementary input tracks could also be coupled to the epigenetic information currently used.

Further research is necessary in the tissue-specific and patient-specific scenario, and also regarding the explainability of the studied features. Whether new predicted cancer genes have cancer driver attributes or not is also left to be explored in follow-up projects such as gene knockout essays or interactome analyses.

Despite these open fronts, the research presented in this work could make its highest impact when designing cancer treatments at a patient-to-patient scale. OriGENE uses data from single patients and compares, at a gene level, the differences between a healthy and a tumor sample. When trained with enough reliable data, the approach we propose would enable the network to discover patient-specific aberrations in the profiles of single genes. This would allow a medical doctor to target the set of aberrantly modified genes, even if they differed between patients diagnosed with the same type of cancer.

Acknowledgments

This work would not have been possible without the help of the following people, who I would like to thank gratefully. I would like to start with Victor, who trusted me and allowed me to pursue this project when success was not granted at all. I would like to thank Mattias, since two words from him were enough to move the project forward. Alin Tomoiaga guided me through my first steps in the world of epigenetics, and this work would not have been born if it were not for him.

Finally, I would like to thank Maria Pilar, Serafí, Arnau, Carla, Gil, Gerard, Roger, Marc, Jake and Marie for their unconditional support and feedback while writing this thesis.

A Data & Code Availability

The original sample files used in our project were produced by Chen K et al. and were part of the paper: Chen K, Chen Z, Wu D, et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* . 2015;47(10):1149-1157., cited as [17]. The samples can be accessed and downloaded from the Gene Expression Omnibus at the following link:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE67471.

The curated sets of genes and the performances of the models with which this work is benchmarked were obtained from Lyu J, Li JJ, Su J, et al. DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. *Sci Adv.* 2020;6(46):eaba6784. Published 2020 Nov 11., cited as [11].

For the tissue-specific prediction, liver-specific CDs were retrieved from Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods.* 2019;16(6):505-507., cited as [22].

All the code can be accessed at the thesis' GitHub repository upon request.

B Methods

B.1 Machine learning methods

Performance metrics. The main performance metrics used in this work are summarized here. The decision threshold or cut for the final node's output was set to 0.5 unless otherwise stated.

- **Sensitivity (Sn).** This metric, also known as recall, measures the number of cancer drivers that a model can pinpoint out of the total number of cancer drivers given.
- **Specificity (Sp).** Magnitude that quantifies how many neutral genes have been correctly classified out of the total number of neutral genes.
- **Precision (Prec).** Percentage of actual cancer drivers detected out of all the predicted positive genes.
- **Accuracy (Acc).** The accuracy of a model quantifies the ratio of correct predictions, including positive and negative genes, out of the total number of genes.
- **Receiver operating characteristic (ROC) curve and Area Under the Curve (AUC).** The ROC curve plots Sn, which coincides with the true positive rate (TPR), against the false-positive discovery rate (FPR = 1-Sp). The AUC is a quantitative estimate of the general performance of a model, without imposing a specific decision threshold. Decision thresholds can be tuned depending on the necessities of the project.
- **Area under the precision-recall curve (AUPRC).** The area below the precision-recall curve (AUPRC) quantifies the average precision of a model. Note that this metric is deeply influenced by the class ratio.

Activation functions. In every node, a weighted summation of the inputs is performed:

$$a = \sum_{k=1}^{K} \omega_k x_k + \omega_0$$

where ω_0 is called a bias term. This value is not directly passed to the next layer, but instead it is used as the argument of a function, the activation function, the output of which will be propagated forward: $\phi(a)$.

The two main activation functions used in this project are the Rectified Linear Unit (ReLU) $\phi(a) = \max(0, a)$, which has strong mathematical and biological motivations and eases the training of deep networks, and the sigmoid $\phi(a) = \frac{1}{1+e^{-a}}$, which is used for the final node in a binary classification scheme.

Loss function. Binary cross-entropy was used for the classification of the final two classes, cancer driver genes vs. neutral genes:

$$E(\vec{\omega}) = -\frac{1}{N}\sum_{n}[d_n \ln(y_n) + (1-d_n)\ln(1-y_n)]$$

The loss function was normalized to the batch size by default.

L2 regularization was added to all the convolutional layers, including the ones inside the Inception modules, which gave an effective loss function like:

$$E(\vec{\omega})' = E(\vec{\omega}) + \alpha\Omega(\vec{\omega})$$

Where the regularization term $\Omega = \frac{1}{2}\sum_i \omega_i^2$ is equivalent to assuming a Gaussian prior for the set of weights $\vec{\omega}$, and the regularization strength was set to $\alpha = 1 \cdot 10^{-2}$.

Dropout, described as the omission of a fraction of nodes randomly chosen when training the network, was introduced for the final layers of the classifier. In order of appearance after the flattening step, the final dropout values were 0.4, 0.3 and 0.1 respectively. Finally, the chosen optimizer was ADAM, with a learning rate set to $4 \cdot 10^{-4}$.

Backpropagation. Given the dependence of the loss function $E(\vec{\omega})$ in the set of weights characterizing the network, this method calculates the gradient of such function with respect to every single weight by means of the chain rule, starting from the last layers and going backwards. Weights will then updated in that direction aiming to minimize the loss.

Ensemble averaging. An ensemble of M networks trained on the same dataset constitutes an added level of regularization. When averaging the outputs of single models, it can be shown that the squared error of the ensemble's predictions is equal to the mean error introduced by each individual network minus a term that quantifies the variance of the different networks in the ensemble. This implies that the performance of an ensemble will be at least as good as the average of the single models constituting it, if not better.

$$y_{ens}(\vec{x}) = \frac{1}{M}\sum_{i=1}^{M} y_i(\vec{x}) \quad ; \quad P_{ens} \geq \frac{1}{M}\sum_{i=1}^{M} P_i$$

Inception modules. Two different 1D INCEPTION modules were used in our final model, which are based on the results presented in [31]. Figure 17, which is part of Figure 5, and Table 5 summarize both modules, since they only differ on the number of filters used. Filter f1 performs merely a kernel 1 convolution, while filters f2 and f3 perform kernel 3 and kernel 5 convolutions after an initial kernel 1 convolution. Finally, filter f4 performs a max pooling on bins of 3 values with stride 1 and a posterior kernel 1 convolution. All the filters are finally stacked on top of each other, since they share the size given an imposed "same" padding.

This configuration of layers is aimed to analyze different features of the signal at varying spatial scales, while keeping the number of weights rather low thanks to the initial kernel 1 convolutions.

Hyperparameter tuning. During the model selection process, the following elements were allowed to vary.

1. Inputs
    - *Sequence length*: $\sim$ [10^4 - 10^5] bp. Shorter genes were zero-padded to the predefined length and larger genes were cropped.
    - *Sequence resolution*: Binning of the signals to bins of [1,10,20,50,100,200] bp.
2. Feature extractor
    - *Number of layers*: $N_l < 10$ (plus pooling layers).
    - *Number of filters per layer*: $N_f \leq 256$.
    - *Type of layers*: Convolutional, MaxPooling, AveragePooling, LSTM, GRU. For the convolutional layers:
        - *Kernel size*: Kernels of size 3 and 5 were tested.
        - *L2 strength,* α: $\alpha = [0, 10^{-1}, 10^{-2}, 10^{-3}]$ were tested.
        - *Stride for the convolutions*: Strides 1 and 2 were allowed.
3. Classifier
    - *Number of layers*: $N_l \leq 5$.
    - *Number of nodes per layer*: $N_n \leq 256$.
    - *Dropout rate*: Dropout values ranging between [0.0-0.8] were allowed.
4. Training parameters
    - *Learning rate*: The ADAM optimizer was tested with learning rates of the orders $[10^{-3}, 10^{-4}, 10^{-5}]$.
    - *Number of epochs*: 50-400 epochs were tested.
    - *Size of the minibatches*: 50-150 paired sequences per minibatch, depending on the number of samples to split into regular parts.
    - *Number of folds,* K: Between 4 and 10 folds, depending on the stage.

The activation function was ReLu for all layers except from the last one, for which the sigmoidal function was used. Also, note that the inception module imposes a 'same' padding that does not apply to the rest of the layers.

B.2 Bioinformatics methods

This section will give an overview of the input data conversion process, from the original SRA files containing the raw sequenced reads to the final bedgraph.sorted files with the continuous H3K4me3 enrichment binned values. A Snakemake workflow was written at a late stage of the project, when assessing the effects of removing duplicated reads on the performance of OriGENE, in order to ease and automatize the process. It can be found in https://github.com/mpielies/MSc-Thesis.

Software required. The software used in this protocol includes:

- **SRA toolkit**: NCBI's SRA toolkit was used to dump the content of the SRA files into fastq files.
- **Bowtie2**: Bowtie2 was used when aligning the reads to the reference human genome hg38.
- **Samtools**: Samtools allowed us to convert the sam files into the corresponding binary format, bam, and sort the reads.
- **Bedtools**: Bedtools was introduced when representing the content of the bam files as enrichment bins in a bedgraph format.
- **Ucsc-bedsort**: Bedsort was useful for sorting the last bedgraph files.

Most of the packages can be installed using *conda* and are included in *bioconda*.

Protocol. The followed protocol with unspecific examples is presented here:

1. Retreive the data:

   ```
   wget https://sra-downloadb.be-md.ncbi.nlm.nih.gov/.../SRR------.-
   ```

2. Dump the SRA files to a fastq file:

   ```
   fastq-dump SRR------.-
   ```

3. Map the file to *hg38*, the latest standarized assembly of the human genome, and get a sam file with all the alignments:

   ```
   bowtie2 -x /.../hg38 -p 4 SRR------.-.fastq -S SRR------.-.sam
   ```

4. Convert the SAM files to their binary equivalent BAM format:

```
samtools view -S -b SRR------.-.sam > SRR------.-.bam
```

5. Sort the resulting BAM files:

```
samtools sort SRR------.-.bam -o SRR------.-.bam.sorted
```

6. Rewrite the content of the BAM files as binned enrichment levels in a bedgraph format:

```
bedtools genomecov -bg -ibam SRR------.-.bam.sorted \
> SRR------.-.bedGraph
```

7. Sort the created bedgraph files:

```
bedSort SRR------.-.bedGraph SRR------.-.bedGraph.sorted
```

   The SRR——.-.bedGraph.sorted resulting file could look like:

   chr1 10000 10007 1

   chr1 10007 10029 2

   chr1 10029 10030 3

   chr1 10030 10031 4

   Where the name of the chromosome appears in the first column, the first and last position of a bin (in basepairs, from left to right in the genome for the positive strand) are shown in second and third columns, and finally the number of reads that were found to overlap in this region.

8. Split the file into chromosomes: The easiest, fastest and most efficient way to split the previous file into chromosomes is by running the following command from the output directory:

```
awk  '{print > "{SRA file title}_"\$1".txt"}' \
.../{SRA file title}.bedgraph.sorted
```

At this point, separate txt files were created for each gene, for example:

GSE67471_GSM1647620_Shifted_2000_H3K4me3_SGK1_neg_OG.txt

Where we have the GEO accession number, the name of the sample, how many basepairs before the TSS were approximately included, the epigenetic markes used, the name of the gene and the strand where it is encoded, and finally its ground truth label. These sequences were further preprocessed as described in Section B.2.1, and were the ones OriGENE worked with.

B.2.1 Sequence preprocessing and normalization

Enrichment profiles and the strand shift. As a consequence of MNase digestion, fragments will come from the DNA at the edge of the nucleosome. Since sequencing takes place only in the 5'-to-3' direction for the 50 first basepairs of each fragment, the reads coming from the positive strand and the negative strand will be mapped on average to slightly shifted locations, which can be seen for the two pillars in Figure 16, with the nucleosome being expected to be found in between. The intensity and definition of the pillars in the original files from [17] arises from to the high number of identical read copies in high enrichment regions, the effects of which are discussed in Appendix C.

Normalization. During the data acquisition steps, factors like the exact amount of genetic material and MNase, the time the enzyme is reacting, etc. unavoidably differ sample to sample, inducing the production of a varying absolute number of fragments and subsequent reads for each experiment.

Aiming to compensate these imbalances, all samples were brought to a shared effective global intensity level by multiplying the corresponding input track by a factor:

$$r = \frac{N_{reference}}{N_{sample}}$$

Where N_{sample} is the number of read counts for each specific sample, presented in Table 4, and $N_{reference}$ was set to 60000 reads in order to change the least the expected read count scale. This procedure could be understood as a total area normalization. Further normalization, for instance gene to gene between different tracks, was discarded because of the following reasons: the area below healthy and matching cancer gene samples is not necessarily expected to be the same, since one could indeed find a shortened profile tail in TSGs for cancer tissue, i.e. less area, or a higher trimethylation level for aberrantly expressed oncogenes, as seen in Section 4.1.1 . A gene specific normalized area could then introduce more bias than the one we want to compensate.

In a normal case scenario, an extra input track obtained from the non-specific precipitation of fragments would have allowed us to assess the relative enrichment of H3K4me3-bound fragments. Since we did not count with such track, relative enrichment levels were

compared directly between healthy and cancer matching samples. Artifacts such as miss-aligned reads leading to fake enrichment bands in repeated regions, e.g., would hence remain in the sequences as an added intrinsical signal noise.

Standardization. To be analyzed by a CNN, all gene samples had to present the same shape: two aligned tracks of 10000 values, one for each basepair (10000,2). To this end, the profiles for negative strand encoded genes were flipped and all genes were cropped or zero-padded depending on their size, which is a common practice, up to the desired length. A window of 10000 basepairs allowed us to capture the enrichment at the promoter of the genes, and how the profile was extended towards the gene body.

Alignment. Given the sample-to-sample dependence of the enrichment bin edges, basepair-resolution enrichment vectors were computed by repeating the value of the left edge of a signal bin until a new intensity value was found. This step permitted the relative alignment of both tracks, but also specifying the position of each new signal value was not required anymore.

In order to avoid zero pooling at the beginning of the signals to fit a predefined grid, we allowed each signal to start at the first data point appearing after 2000 bp before the beginning of the gene, and then cropped the start of the less restrictive track. This means that sequences did not start exactly 2000 bp before the TSS, but that was not a strong requirement given that CNNs are meant to handle translated patterns, and also the promoter will display different features in different genes: we need to frame the enrichment profile, and 1000-2000 bp are the standard values used in the literature.

The locations of the genes in the genome were obtained from [33].

C Duplicated Read Removal

Another matter of concern was the nature of identically duplicated reads and their impact in the performances of the models. Possibly due to the 10 PCR amplification cycles stated in the library preparation protocol, a considerable number of reads appeared to have exactly the same coordinates, especially in high enrichment regions[15]. Even though some of them could have a natural origin, whether these duplicates contributed to the information content of the signals or were in fact artifacts hiding the relevant structure needed some clarification.

Despite the serious modifications introduced in the H3K4me3 enrichment profiles, OriGENE performed rather similarly when trained and validated with re-normalized sequences where all duplicates were removed, as seen in Table 6 and Figure 18. Accuracies were indeed slightly higher than in the original validation, mainly due to a rise in specificity in three of the samples, but there was no clear gain regarding the ability of the models to pinpoint actual cancer drivers. Thus, the magnitude of the observed performance improvements did not justify introducing such a strong constraint, and therefore we decided to proceed with the signals as they were published.

[15]More than half of the reads for Liver II, e.g., were non-unique. For more information visit https://github.com/mpielies/MSc-Thesis

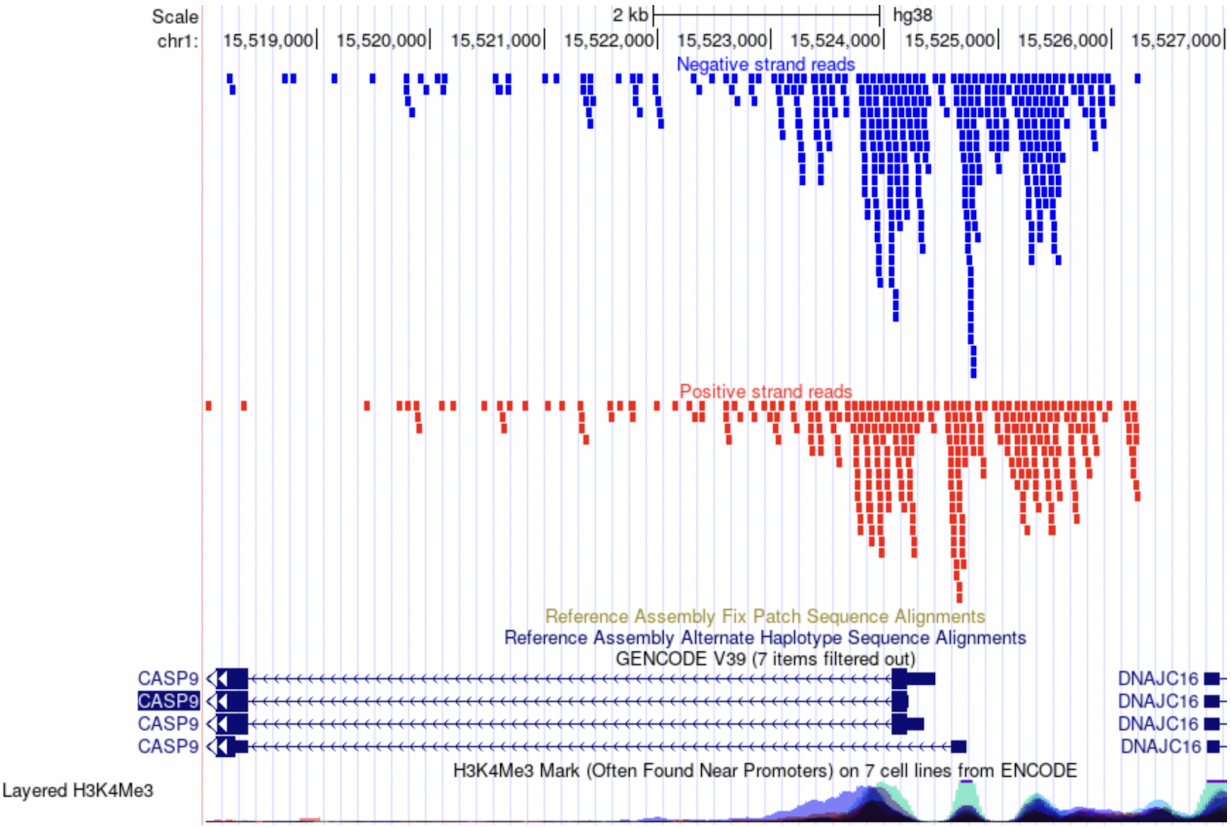


Figure 16: **Visualizing the H3K4me3 enrichment profiles for CASP9.** The promoter region and beginning of CASP9, encoded in the reverse strand, are shown. The unduplicated 50 bp long reads mapped to hg38, plotted as red and blue rectangles, are piled up. The binned enrichment profile emerging as a combination of the reads from both strands leads to the signals shown in Figure 4, which are comparable to the profiles shown below for 7 different cell lines when flipped, with two clear H3K4me3 devoided regions. The distance between the expected position of the reads coming from both strands can be appreciated, around the central peak corresponding to a well located nucleosome, as two shifted pillars.

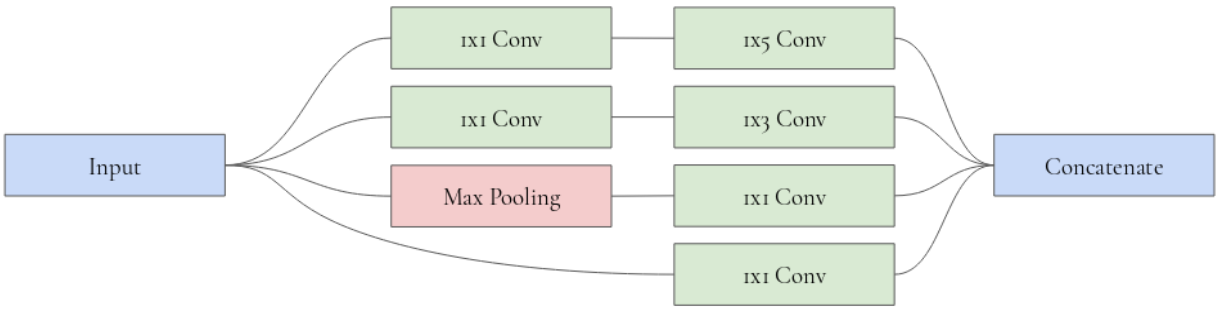


Figure 17: **INCEPTION modules.** 1D adaptation of the Inception module introduced in [31]. The outputs of all filters are stacked keeping the spatial information for the subsequent layer to compare them.

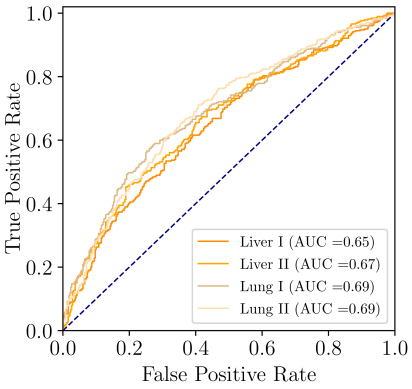


Figure 18: **ROC curves for the validation sets of sequences with all duplicates removed**

TABLE IV. Dataset information

NAME	HEALTH STATUS	GEO ACCESSION NUMBER	READS
Lung I	*Healthy tissue*	GSM1647618	59251101
	Cancer tissue	GSM1647619	75836608
Liver I	*Healthy tissue*	GSM1647620	62763610
	Cancer tissue	GSM1647621	65673468
Liver II	*Healthy tissue*	GSM1647622	64811514
	Cancer tissue	GSM1647623	61400238
Lung II	*Healthy tissue*	GSM1647624	34545390
	Cancer tissue	GSM1647625	32711856

Table 4: **Dataset information**

TABLE V. Number of filters for the INCEPTION module

INCEPTION MODULE	F1	F2		F3		F4
	Output	*Input*	*Output*	*Input*	*Output*	*Output*
A	8	8	8	8	8	8
B	4	4	4	4	4	4

Table 5: **Number of filters for the INCEPTION modules A and B.**

TABLE VI. Performance metrics when removing duplicated reads

STAGE	TISSUE	ACC	SP	SN	PREC	AUC	90% CI
Validation Duplicates Removed	*Liver I*	0.609	0.626	0.592	0.613	0.648	[0.615 - 0.682]
	Liver II	0.621	0.721	0.521	0.651	0.666	[0.632 - 0.700]
	Lung I	0.657	0.723	0.584	0.683	0.685	[0.655 - 0.718]
	Lung II	0.649	0.684	0.613	0.660	0.688	[0.656 - 0.721]

Table 6: **Performance metrics of OriGENE when trained and tested using sequences with all duplicates removed.**

References

[1] Gonzalez-Perez A, Lopez-Bigas N. Functional impact bias reveals cancer drivers. *Nucleic Acids Res.* 2012;40(21):e169.

[2] Dees ND, Zhang Q, Kandoth C, et al. MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* 2012;22(8):1589-1598.

[3] Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature.* 2013;499(7457):214-218.

[4] Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics.* 2013;29(18):2238-2244.

[5] Davoli T, Xu AW, Mengwasser KE, et al. Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell.* 2013;155(4):948-962.

[6] Reimand J, Bader GD. Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol Syst Biol.* 2013;9:637.

[7] Tokheim CJ, Papadopoulos N, Kinzler KW, Vogelstein B, Karchin R. Evaluating the evaluation of cancer driver genes. *Proc Natl Acad Sci U S A.* 2016;113(50):14330-14335.

[8] Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. *Genome Biol.* 2016;17(1):128. Published 2016 Jun 16.

[9] Dietlein F, Weghorn D, Taylor-Weiner A, et al. Identification of cancer driver genes based on nucleotide context. *Nat Genet.* 2020;52(2):208-218.

[10] Chandrashekar P, Ahmadinejad N, Wang J, et al. Somatic selection distinguishes oncogenes and tumor suppressor genes. *Bioinformatics.* 2020;36(6):1712-1717.

[11] Lyu J, Li JJ, Su J, et al. DORGE: Discovery of Oncogenes and tumoR suppressor genes using Genetic and Epigenetic features. *Sci Adv.* 2020;6(46):eaba6784. Published 2020 Nov 11.

[12] Schulte-Sasse R, Budach S, Hnisz D, Marsico A. Integration of multiomics data with graph convolutional networks to identify new cancer genes and their associated molecular mechanisms. *Nat. Mach. Intell.* 3 (2021): 513-526.

[13] Wu CT, Morris JR. Genes, genetics, and epigenetics: a correspondence. *Science.* 293.5532 (2001): 1103-1105.

[14] Barski A, Cuddapah S, Cui K, et al. High-resolution profiling of histone methylations in the human genome. *Cell*. 2007;129(4):823-837.

[15] Wysocka J, Swigut T, Xiao H, et al. A PHD finger of NURF couples histone H3 lysine 4 trimethylation with chromatin remodelling. *Nature*. 2006;442(7098):86-90.

[16] Jenuwein T, Allis CD. Translating the histone code. *Science*. 2001;293(5532):1074-1080.

[17] Chen K, Chen Z, Wu D, et al. Broad H3K4me3 is associated with increased transcription elongation and enhancer activity at tumor-suppressor genes. *Nat Genet* . 2015;47(10):1149-1157.

[18] Zhao D, Zhang L, Zhang M, et al. Broad genic repression domains signify enhanced silencing of oncogenes. *Nat Commun*. 2020;11(1):5560. Published 2020 Nov 3.

[19] Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, et al. Integrative analysis of 111 reference human epigenomes. *Nature*. 2015;518(7539):317-330.

[20] Xia B, Zhao D, Wang G, et al. Machine learning uncovers cell identity regulator by histone code. *Nat Commun* . 2020;11(1):2696. Published 2020 Jun 1.

[21] Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers. *Nat Rev Cancer*. 2018;18(11):696-705.

[22] Lever J, Zhao EY, Grewal J, Jones MR, Jones SJM. CancerMine: a literature-mined resource for drivers, oncogenes and tumor suppressors in cancer. *Nat Methods*. 2019;16(6):505-507.

[23] Edgar R, Domrachev M, Lash AE. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*. 2002;30(1):207-210.

[24] Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 A resolution. *Nature*. 1997;389(6648):251-260.

[25] Deribe YL, Pawson T, Dikic I. Post-translational modifications in signal integration. *Nat Struct Mol Biol*. 2010;17(6):666-672.

[26] Alabert C, Groth A. Chromatin replication and epigenome maintenance. *Nature Reviews. Molecular Cell Biology*.2012 Mar;13(3):153-67.

[27] Santos-Rosa H, Schneider R, Bannister AJ, et al. Active genes are trimethylated at K4 of histone H3. *Nature*. 2002;419(6905):407-411.

[28] Suvà ML, Riggi N, Bernstein BE. Epigenetic reprogramming in cancer. *Science*. 2013;339(6127):1567-1570.

[29] Davis IJ, Pattenden SG, Chapter 1-3 - Chromatin Accessibility as a Strategy to Detect Changes Associated With Development, Disease, and Exposure and Susceptibility to Chemical Toxins, *Toxicoepigenetics, Academic Press*, 2019, Pages 85-103, ISBN 9780128124338.

[30] Schones DE, Cui K, Cuddapah S, et al. Dynamic regulation of nucleosome positioning in the human genome. *Cell*. 2008;132(5):887-898.

[31] Szegedy C, Liu W, Jia Y, et al. Going Deeper with Convolutions. *CVPR*, 2015.

[32] Baylin SB. DNA methylation and gene silencing in cancer. *Nat Clin Pract Oncol*. 2005;2 Suppl 1:S4-S11.

[33] Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. *Genome Res*. 2002;12(6):996-1006.