Master's Thesis in Chemical Biotechnology

Predicting Interactions of RNA-Binding Proteins Regulated by N(6)-Methyladenosine Modification

Sofia Martello

Master´s Thesis in Chemical Biotechnology

Predicting Interactions of RNA-Binding Proteins Regulated by N(6)-Methyladenosine Modification

Author:	Sofia Martello
Advisor:	Giulia Cantini
External Supervisor:	Prof. Dr. Annalisa Marsico
Internal Supervisor:	Prof. Dr. Dominik Grimm
Submission Date:	30th June 2023

I confirm that this master's thesis in chemical biotechnology is my own work and I have documented all sources and material used.

Straubing, 30th June 2023 Sofia Martello

Acknowledgments

I would like to express my deepest gratitude to Prof. Dr. Annalisa Marsico for providing me with the incredible opportunity to work in her esteemed research team, "Computational RNA Biology", at Helmholtz Munich. Her guidance, mentorship, and support have been invaluable throughout my journey.

I extend my heartfelt thanks to Giulia Cantini for her constant advice and unwavering assistance throughout the entire duration of this project. Her expertise and encouragement have been instrumental in shaping the direction and outcomes of my research.

I am profoundly grateful to my family, my parents Ivo and Valentina, and my siblings Iris, Sara, and Pietro, who have consistently been a source of inspiration and a pillar of strength. Their relentless belief in me has been a driving force in my pursuit of excellence.

I would like to express my gratitude for the friendships I have cultivated during my time at TUMCS. In particular, I want to extend my sincere appreciation to Ian, Teresa, Stefano, and Valeria for their unwavering support, attentive listening, and invaluable suggestions. Their presence has truly enriched my life and made it extraordinary.

Lastly, I would like to acknowledge myself and my persistent determination. This journey would not have been possible without my resilience, perseverance, and unyielding commitment to my goals.

Thank you all for your invaluable contributions to my academic and personal growth.

Abstract

RNA molecules play diverse roles in cellular processes and gene regulation. The binding of RNA molecules to proteins, known as RNA-protein interactions, is a crucial event that regulates various mechanisms of RNA processing. This study focuses on RNA-binding proteins (RBPs) and investigates how their affinity with RNA can be affected by modifications. One prominent methylation is N(6)-methyladenosine (m6A), the most abundant internal modification in eukaryotic mRNA, which has been implicated in post-transcriptional gene regulation. This project aims to understand the impact of m6A modifications on RNA-protein interactions and develop models that integrate RNA sequence and m6A data to predict the binding preferences of RBPs. Deep learning techniques, specifically convolutional neural networks (CNN), are employed to capture complex patterns and develop accurate models. The integration of m6A data into the models allows for an attempt to understand the interplay between m6A and RBPs in RNA-protein interactions. The study utilizes single nucleotide resolution CLIP data to train a binary classifier. Multiple datasets are formed, including positive sets of sequences bound to specific RBPs, negative sets of unbound sequences, and negative sets of sequences bound to other RBPs. With the incorporation of these datasets, the model's performance in discriminating between bound and unbound sequences is evaluated. The results demonstrate promising outcomes in classifying bound and unbound sequences, providing insights into the distinctive features recognized by the model for RBP binding. Interestingly, the findings suggest that m6A modifications may significantly influence the binding preferences of RBPs. The observed correlation between m6A modifications and the binding affinity of RBPs supports the notion that m6A plays a crucial role in shaping RNA-protein interactions. The concept of the "m6A-switch" mechanism, where m6A alters the RNA three-dimensional structure and enhances the affinity for RBP binding, has been supported by the results of this research.

Overall, this study contributes to our understanding of the interplay between m6A modifications and RBPs in RNA-protein interactions. The developed models and findings lay the groundwork for further investigations in this field, including exploring the impact of other RNA modifications and incorporating additional features into the models. This research advances our knowledge of RNA biology and opens new avenues for future studies in the field of RNA-protein interactions.

Contents

List of Figures

List of Tables

1. Introduction

RNA molecules perform essential functions in cells, namely by acting as messengers between DNA and proteins (messenger RNA), facilitating protein synthesis (ribosomal RNA and transport RNA permit indeed translation), regulating gene expression (case of microRNA and long non-coding RNA), and catalyzing chemical reactions (ribozymes act similarly to enzymes). They contribute to the complexity of cellular processes and play a crucial role in gene expression regulation and protein production [1, 2]. The binding of RNA molecules to proteins is a key event that can regulate various mechanisms of RNA processing: the focus of this project is on RNA-binding proteins (abbreviated as RBPs from here on) and how their affinity with the RNA could be affected by modifications. RNA modifications play a critical role in post-transcriptional gene regulation, particularly by influencing RNA stability, splicing, localization, and translation [3]. One prominent modification is N(6)-methyladenosine (m6A), which is the most abundant internal modification in eukaryotic mRNA. m6A modification has been implicated in several key aspects of RNA biology, including mRNA decay, alternative splicing, RNA structure, translation efficiency, and RNA-protein interactions. Understanding the impact of m6A methylation on RNA-binding proteins and on their affinity for RNA can provide valuable insights into the regulatory mechanisms governing RNA processing and gene expression.

RNA-binding proteins, similarly, have various roles in the regulation of gene expression at the post-transcriptional level, being essential in the process of alternative splicing, polyadenylation, translation, stabilisation, and localization of transcripts. They are also implicated in the formation of the ribonucleoprotein (RNP) complex and regulate RNA turnover, quality control, and transport [4]. Any dysregulation of said processes can lead to various diseases: numerous studies have indeed linked atypical expression or mutation of RBPs to various medical conditions such as neurodegenerative, metabolic diseases, and cancer [5, 6, 7, 8]. For instance, mutations in RBP genes have been found to provoke spinal muscular atrophy and amyotrophic lateral sclerosis, two diseases that affect the motor neurons and the nervous system [9]. Moreover, the identification of RBP binding sites is relevant for the annotations of genes, since it delivers information about functional regions in the genome, for the prediction of RNA structure, which is essential for the understanding of its function and stability, and eventually for evolutionary analysis. In fact, the identification of conserved RBP binding sites across species could help provide insights into RNA regulatory mechanisms [6, 10].

Deep learning techniques are helpful to capture complex patterns and develop accurate models in multiple situations. However, it is difficult to accurately model both the dynamic nature of RNA-protein interactions in vivo and the effects of RNA modifications on these interactions. As a matter of fact, RBPs are not static entities; they can interact with other cellular components and undertake conformational changes that are laborious to be modelled. Besides, the binding of RNA-binding proteins to the RNA strand is influenced by a multitude of factors: RNA secondary and tertiary structure, various modifications, the protein concentration, and other factors that depend on the cellular context. For instance, RNA modifications such as N(6)-methyladenosine have been shown to alter RNA-protein interactions by modifying the 3D structure of RNA [11]. Therefore, predicting the binding of RNA-binding proteins to RNA in vivo is an intricate challenge that necessitates the consideration of cellular dynamics. One crucial aspect to investigate is the impact of m6A on RNA-protein interactions, as this methylation has the potential to influence or serve as a predictive factor for binding. To address this, it is essential to develop integrated models that incorporate both RNA sequence and modification data, enabling a comprehensive understanding of the interplay between m6A and RBPs in RNA-protein interactions.

After these considerations and the input from other works [12] stating that there is a relationship between m6A methylation and protein binding, it has been reasoned that an investigation in this direction is imperative. The aim of this project is to explore the effects of m6A on RNA-protein interactions and develop models that integrate RNA sequence and m6A data to predict the binding preferences of RBPs in vivo and study the impact of m6A modifications on the regulatory code that controls binding. To research this topic, a binary classifier based on convolutional neural networks (CNN) has been designed and single nucleotide resolution CLIP data, representing both RBP binding sites and m6A sites, have been exploited to train it. The integration of binding site information and m6A site data involved the formation of three distinct sets. The positive labelled set consists of sequences bound to a specific RNA-binding protein, while the first negative labelled set comprises sequences that are not bound to any RBP. Additionally, the second negative labelled set is composed of sequences that are not bound to the specific RBP of interest, but rather to other RBPs. The utilization of multiple datasets provides an opportunity to compare results and gain insights into the performance of the model in effectively discriminating between bound and unbound sequences [13]. With the integration of m6A information, this study aims to explore whether there exists a correlation between the binding affinity of RNA-binding proteins and the presence of m6A modifications in the vicinity of the binding sites.

The manuscript is structured as follows. The initial section, "Concepts," provides a comprehensive overview of the fundamental background covering pertinent biological and computational concepts. Subsequently, the "Materials and Methods" section elucidates the specific datasets employed in the study, outlining the preprocessing techniques implemented, and delineating the architecture of the proposed model. Following this, the "Results" section presents an exhaustive analysis and presentation of the obtained outcomes, categorized by specific cases. The section "Discussion" incorporates a comprehensive discussion of the findings. Lastly, the manuscript culminates with the "Conclusion" section, summarizing the key findings and implications drawn from the study’s results.

2. Concepts

2.1. Biological Background

2.1.1. RNA-binding Proteins

Biological sequences organize life on earth: these are DNA (deoxyribonucleic acid), RNA (ribonucleic acid) and proteins. Although the DNA sequences of several cells of the same organism are identical (if mutations are ignored), the set of RNA and proteins available at different time points is independent and defines the specific functions of the cell [14]. The structure and function of DNA and RNA are closely related. DNA is composed of four nitrogenous bases: adenine (A), guanine (G), cytosine (C), and thymine (T). These bases are arranged in a specific order or sequence that determines the genetic code. The base sequence of DNA is usually complementary to the base sequence of RNA, which consists of the same four bases, except for thymine, which is replaced by uracil (U). During transcription, the genetic information encoded in DNA is transcribed into RNA, which serves as a template for protein synthesis. DNA and RNA also differ in structure, in fact DNA is a double-stranded helix, whereas RNA is generally single-stranded. This structural difference allows RNA to fold into complex structures and perform a variety of functions, including catalyzing chemical reactions and regulating gene expression, roles that are typically associated with proteins [15].

Proteins that play a crucial role in the regulation of gene expression are called transcription factors (TFs). These polymers are capable of binding DNA, controlling the rate at which particular genes are transcribed into RNA, activating or repressing gene transcription. TFs do not bind to the gene locus itself but to so-called regulatory regions, namely promoters and enhancers, which are involved in the initiation of transcription [16, 17].

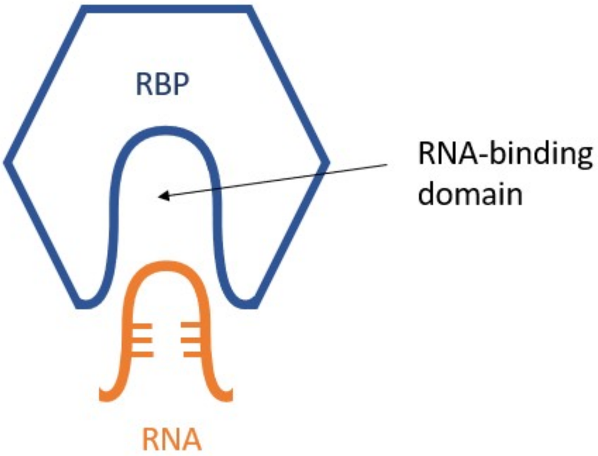


Figure 2.1.: An RNA-binding protein (RBP) modulates RNA processing and functionality by binding RNA molecules via specific RNA-binding domains.

Similar to TFs that bind to DNA, RNA binding proteins are proteins that can physically interact with RNA [18], see Figure 2.1. RBPs can be classified into several families based on their structural and functional properties. Each family of RBPs interacts with RNA in different ways, and different RBPs can bind to independent RNA sequences or structures, granting a wide range of regulatory functions [19, 20]. Proteins that bind RNA usually require an RNA-binding domain, referred to as a binding motif, but some RBPs can bind RNA via an intrinsically disordered region. RBPs generally recognise shorter RNA sequences than those recognized by TFs; usually around 3-5 bases in length. Similar to TFs, RBPs also use mechanisms to ensure certain bindings, such as multiple RNA-binding domains, cooperative binding of multiple proteins, competitive binding, and preference for specific RNA secondary structures [21]. RBPs are involved in many biological processes, including alternative splicing, mRNA stability, and translation. Alternative splicing,i.e. the process by which different exons of pre-mRNA combine to generate multiple protein isoforms or variations, is controlled by RBPs that bind to specific splice sites and regulate the splicing machinery [22]. mRNA stability is also regulated by RBPs, which bind to specific sequences in mRNA and promote or inhibit its degradation, for example by adding a capping to the RNA sequence [23, 24]. Finally, RBPs can also regulate translation by binding to the 5′ and 3′ untranslated regions of mRNA, regulating access to ribosomes, and affecting translation efficiency [25] (Figure 2.2).

2.1.2. RNA Modifications

RNA undergoes post-transcriptional modifications that contribute to the complexity of cellular and biological processes, similar to DNA. These modifications can be reversible or irreversible and play important roles in various cellular processes, including transcription,

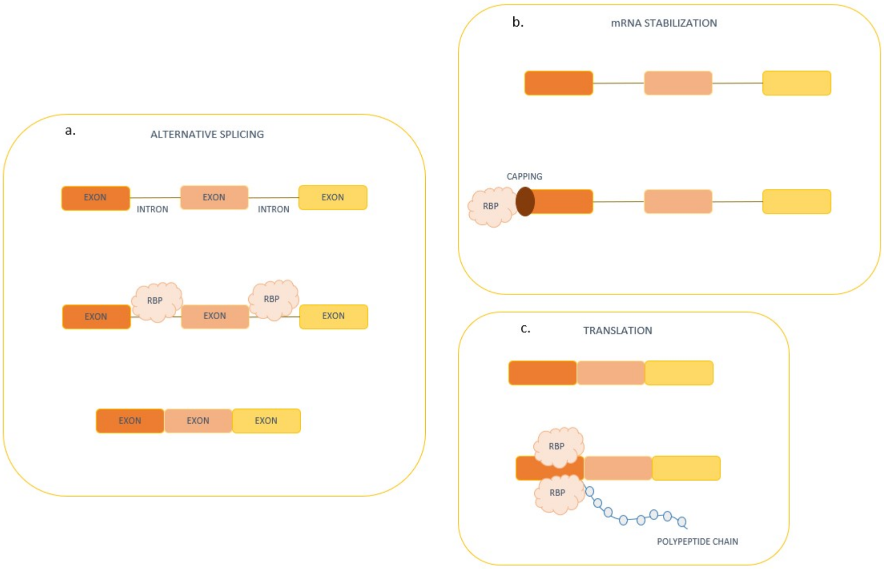


Figure 2.2.: Examples of RBP functions in biological processes: a. Alternative Splicing, b. mRNA stability, c. Translation regulation

pre-mRNA splicing, RNA export, mRNA translation, and RNA degradation [3]. Such modifications collectively shape the cellular transcriptome and proteome [26, 27, 28].

There are several types of RNA modifications, including but not limited to methylation (the addition of a methyl group), acetylation (the addition of an acetyl group), pseudouridylation (isomerization of uridine to pseudouridine), poly(A) tail addition (the addition of adenine nucleotides at the 3′ end of mRNA), and A-to-I editing (conversion of adenosine to inosine). These modifications can occur at different regions of RNA, such as the base, sugar, or phosphate backbone, and are catalyzed by specific enzymes. Each modification can have varying effects on RNA function, and their combinations further promote the intricacy of RNA regulation [29]. Various sorts of modifications, such as methylation, pseudouridylation, and base modifications, can alter the chemical composition and physical properties of RNA, which can lead to changes in RNA structure by affecting base pairing, stability, flexibility, and interactions with other molecules. Dysregulation of RNA modifications has been associated with various diseases, including cancer and neurological disorders, by having an impact RNA processing and gene expression [27].

In this project, the focus is on N6-methyladenosine (m6A), the most prevalent internal modification in higher eukaryotes, present in varying percentages across different species and tissues [30]: m6A involves the addition of a methyl group to the N6 site of the adenosine base. Although initially identified in the 1970s in poly(A) and RNA, its recognition as a predominant component of mRNA came later, through transcriptome-wide mapping studies, revealing its presence in thousands of transcripts, in regions near the stop codon, coding sequence, 3′ UTR, and 5′ UTR [31]. The presence of m6A can induce conformational changes, for instance with the alteration of base stacking, hydrogen bonding, and RNA secondary structure. These structural alterations can subsequently influence RNA-protein interactions, including the binding of RNA-binding proteins and other regulatory factors [11].

The discovery of m6A as a major mRNA modification has sparked significant interest in the field of RNA biology, leading to investigations into its role in various cellular processes and its impact on different diseases. Recent studies have confirmed that m6A serves as a central regulator of mRNA stability, promoting the decay of unstable mRNAs. Additionally, m6A has been shown to play a role in alternative splicing regulation and translation initiation, facilitating the translation of specific mRNAs. Understanding the mechanisms and functions of m6A and other RNA modifications holds significant implications for disease research and treatment [32].

2.1.3. Data Production

Various protocols are available for the collection of RNA and RBP data. For the analysis of the transcriptome, i.e. the entire set of RNA transcripts produced by a cell or a tissue at a specific time point, RNA-seq is the state-of-the-art technology available. RNA-seq is a next-generation sequencing-based technique that includes converting RNA into cDNA and using high-throughput sequencing technologies to quantify all RNA species present in a sample, such as mRNA, non-coding RNA, and splice variants [33]. Compared with traditional methods like microarrays, RNA-seq has a number of advantages: higher sensitivity and accuracy, and the detection of new transcripts and variants since it does not require prior knowledge of the investigated genome [34]. For this reason, it is suitable for the investigation of non-model organisms or poorly annotated genomes.

RNA-protein interactions can be studied using the CLIP-seq protocol(Cross-Linking and Immunoprecipitation), which has transformed the field by enabling researchers to identify these binding sites on a transcriptome-wide scale [35]. This method involves cross-linking RNA and proteins, followed by immunoprecipitation and sequencing of the RNA molecules to identify protein binding sites, see Figure 2.3. However, the CLIP method has limitations in sensitivity, specificity, and produces technical artefacts [36]. Nevertheless, various modifications of the CLIP technique have emerged, namely iCLIP [37] (individual-nucleotide resolution CLIP), eCLIP [38] (enhanced CLIP), and irCLIP [39] (in vitro selection and sequencing of RNA-protein interactions).

The eCLIP method is considered to have higher sensitivity and specificity compared to other methods, and it can detect additional binding sites that were missed by other methods [38]. The eCLIP method implements an efficient experimental workflow that includes several key alterations to the standard CLIP protocol. One variation is the incorporation of 4-thiouridine (4sU) into nascent RNA to mark newly synthesised transcripts, which enhances the signal-to-noise ratio. Another adjustment is a modified ligation step that reduces adapter dimers. Additionally, eCLIP employs an original computational pipeline that accounts for PCR duplicates, filters out low-quality reads, and normalizes the read coverage. This method has been shown to be highly effective in identifying RNA-protein interactions and has been widely utilized in the field [36].

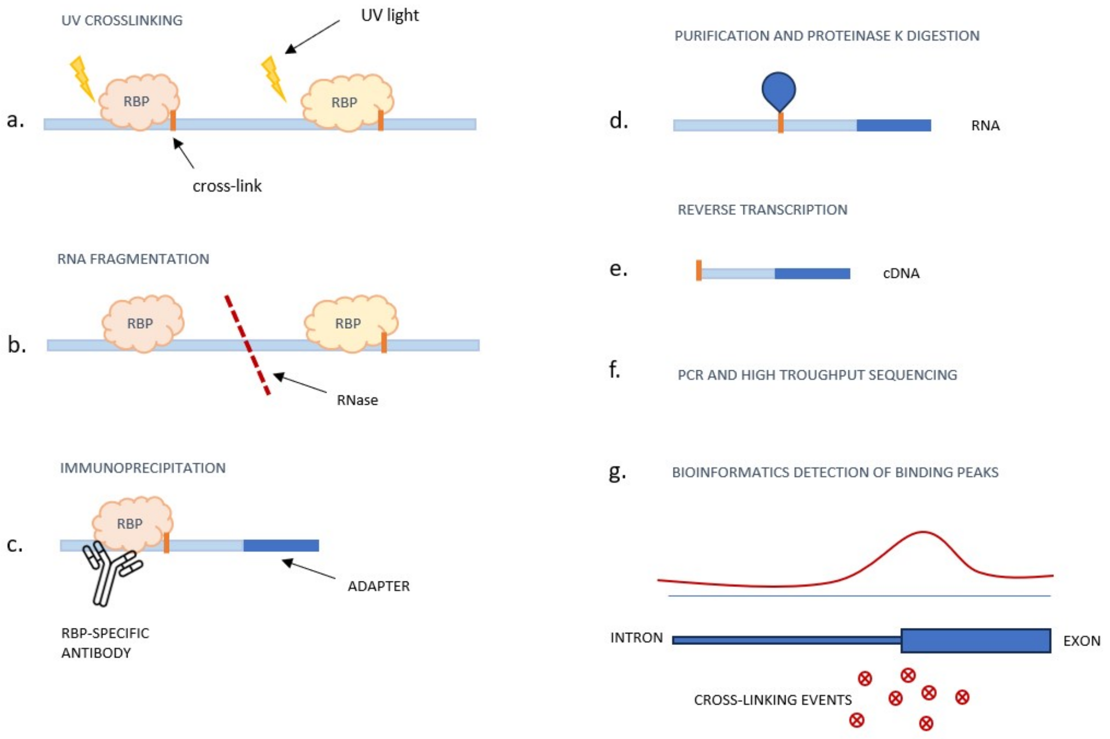


Figure 2.3.: Overview of the key stages in the cross-linking and immunoprecipitation (CLIP) protocol.

(a.-b.) Generation of cross-linked lysate. (c.-d.) Isolation of targeted cross-linked RNA fragments. (e.-g.) Sequencing and analysis of cDNA library.

Another decisive point for this project is the collection of m6A site data. One of the methodologies that allow this is miCLIP [31], which applies UV light-induced crosslinking of RNA to m6A antibodies. Subsequently, reverse transcription of the crosslinked RNA, coupled with the aforementioned m6A antibodies, leads to the generation of a distinct pattern of mutations or truncations in the resulting cDNA. This, once sequenced and analysed, enables precise identification of m6A residues. This protocol has been shown to provide single-nucleotide maps of m6A modification sites in various RNA targets, including mRNA and small non-coding RNA, with high specificity and sensitivity. Moreover, miCLIP has shown compatibility with low-input RNA samples and has several advantages over previous methods for mapping m6A sites.

Nevertheless, there are still various problems connected to this and more methods exploiting antibodies, some of them being PA-seq [40], m6A-seq2 [41] and m6A-LAIC-seq [42]. The resolution of antibody-based methods is limited, and they lack quantitative information for specific modification sites; for example, the specificity of m6A identification can be compromised by the promiscuous binding of m6A antibodies to certain RNA sequences or structures. On that account, antibody-free methods have been developed, for example glyoxal and nitrite-mediated deamination of unmethylated adenosines (GLORI) [43] keeps intact just the adenosines that are m6A methylated, modifying the rest of them, to obtain clear identification of m6A sites.

2.2. Computational Background

2.2.1. Peak Calling

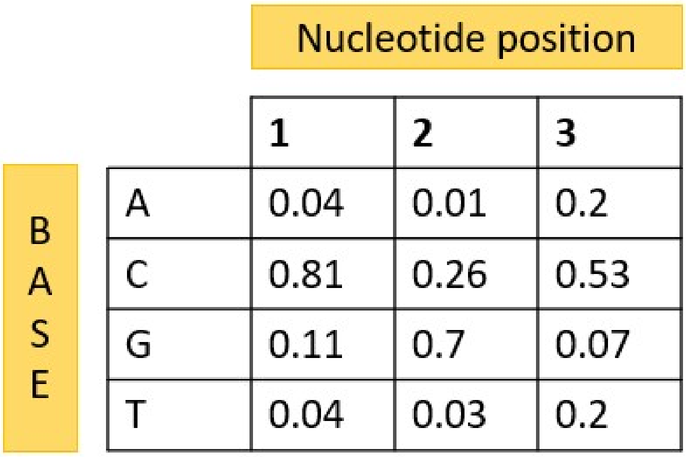
Peak calling is a fundamental computational approach employed to identify regions of the genome that exhibit enriched signals in high-throughput sequencing data, specifically for CLIP data. This technique plays a crucial role in understanding the underlying biological processes involving protein-RNA interactions [44]. Various peak calling algorithms leverage the coverage features of CLIP-seq and control data to detect genomic regions displaying presence of protein binding. These methods require the alignment of sequencing reads and generate a list of regions, known as "peaks," which indicate potential sites of protein-RNA interaction. Each detected peak is assigned a significance score, providing a quantitative measure of the level of enrichment observed in the experimental data compared to the control background [45]. The process of peak calling is indispensable in the analysis of CLIP data as it enables researchers to identify specific genomic regions where RNA molecules are potentially interacting with RNA-binding proteins (RBPs).

In general, peak calling requires four steps: (1) preprocessing of the data, (2) peak detection, (3) peak annotation, and (4) statistical analysis [46, 47]. The first step consists of cleaning and filtering raw sequencing data to eliminate low-quality reads, adaptor contamination, and PCR duplicates. This step is important to ensure accurate peak detection and reduce false positives. The next step is to identify sections of the genome that have considerably higher read counts than the background level. Peak detection techniques range from simple threshold-based approaches, where a predefined threshold is set for read counts, to more advanced algorithms that utilize statistical models to account for fluctuation and noise in the data. After identifying peaks, they must be annotated with functional information, such as their position relative to gene features, transcription factor binding sites, or histone modifications. This stage is crucial for recognising the biological relevance of the detected peaks and developing theories about the underlying molecular pathways. Finally, peak calling concludes with statistical analysis to identify the importance of determined peaks and to compare peak profiles across different experimental conditions. Performing differential peak analysis to discover peaks that are considerably differentially enriched between various groups or time periods, or clustering analysis to identify groups of peaks with similar characteristics, are examples of statistical analysis.

Examples of widely used peak callers for CLIP data include CLIPper, Piranha, and PureCLIP. CLIPper [38] is a strand-specific peak calling tool developed by ENCODE. It incorporates advanced statistical models and algorithms to effectively handle data noise and fluctuations. To ensure confidence in the identified binding sites, CLIPper provides various measures for quality control, such as assessing the number of unique reads and the percentage of reads within a peak. Researchers regularly rely on CLIPper for its effectiveness and valuable contributions to the field [48]. Notably, CLIPper serves as the underlying ENCODE peak caller for the binding sites used in this study. Piranha [49] is another widely used tool that performs strand-specific peak calling. While it accepts variables, it does not explicitly correct for non-specific background signals. Piranha calculates a genome-wide significance threshold by modelling the distribution of read counts on a binwise basis. When this threshold is exceeded, a peak is called. It is important to note that both Piranha and CLIPper are peak calling methods that are insensitive to individual crosslink sites. PureCLIP [48] is specifically designed to extract target-specific protein-RNA interaction marks from iCLIP/eCLIP-seq data. By considering regions enriched with protein-binding fragments and incorporating iCLIP/eCLIP-specific cleavage patterns, PureCLIP calls individual crosslinks. It takes advantage of a non-uniform hidden Markov model (HMM) and incorporates additional elements into the model, such as a non-specific background signal from input experiments, to reduce the number of false positives. As a result, PureCLIP demonstrates high specificity in detecting crosslinking sites. Researchers often evaluate and compare different peak callers to select the most suitable one for their specific needs. The choice of a peak caller depends on the research question, the characteristics of the CLIP data, and the desired analysis outcomes. By considering these factors, researchers can leverage the strengths of each peak caller to enhance their understanding of protein-RNA interactions in CLIP data analysis.

2.2.2. Traditional Computational Methods

For many years, sequence motifs have been the preferred technique for inferring from sequence data because they provide information about protein binding preferences and may be used to evaluate the impact of mutations that modify those binding preferences [50]. Sequence motifs are patterns that transcription factors (TFs) and RNA-binding proteins (RBPs) can recognize. These patterns can vary by one or more bases, and it is observed that certain sites change very rarely, while others modify frequently [21]. To facilitate the visual evaluation of a collection of sequences, sequence motifs are represented by stacking the possible bases of each position on top of each other. The heights of these stacks are proportional to their observed frequencies, which are collected in a position frequency matrix (PFM) - a matrix of size m x n, where m represents the number of possible bases and n represents the length of the sequences [51] (Figure 2.4).



Nucleotide position

BASE

	1	2	3
A	0.04	0.01	0.2
C	0.81	0.26	0.53
G	0.11	0.7	0.07
T	0.04	0.03	0.2

Figure 2.4.: Example of Position Frequency Matrix.

The PFM represents the frequencies of each base at different positions. The sum of base frequencies for each position adds up to 1.

Experimentally derived sequences that are assumed to share common motifs are typically long (hundreds of bases), but the starting positions of motifs and the motifs themselves are unidentified. Therefore, de novo motif-finding tools capable of discovering overrepresented motifs in an unsupervised fashion have been developed. Many tools and methods have been presented in the past [52], with MEME [53] being one of the most frequently utilized. MEME identifies motifs by maximizing the information content of position frequency matrices using an expectation-maximization approach. Following the de novo discovery of motifs, it is typically interesting to computationally predict further motif occurrences by scanning sequences that were not experimentally covered with a position frequency matrix. To score each position in a sequence of interest based on its resemblance to an established motif, RSAT [54] evaluates not only the sequence of interest but also background sequences. This evaluation allows for the calculation of a score threshold that permits the detection of statistically significant motif hits.

In addition to the motif finding and scanning approach, it is possible to create a pipeline for binary classification of sequences based on whether or not they include a motif hit. This pipeline combines motif discovery and scanning to predict, for example, protein binding in sequences that have not been previously studied. However, dedicated supervised machine learning methods that incorporate feature extraction and scoring into a single model have been shown to outperform the motif finding/scanning approach for classification tasks like protein binding prediction [55].

2.2.3. Deep Learning Methods for RNA binding sites prediction

Artificial neural networks (ANNs) are machine learning algorithms, designed to imitate the function and structure of the human brain. ANNs are structures comprised of linked nodes that process data and generate output based on such information. They have grown in prominence in recent years because of their capacity to learn and adapt to new data, making them useful tools in various applications such as image recognition, speech recognition, and natural language processing. The artificial neuron, often known as a node, is the fundamental unit of an ANN. The node accepts input from other nodes or external sources and generates output in response. The output is subsequently sent to other network nodes. Each node in the network is linked to the others via a series of weighted connections. Weights reflect the strength of the link between nodes and are modified throughout the learning process to improve network performance [56].

A supervised learning procedure is used to train ANNs, in which the network is given a collection of labelled samples and instructed to understand the relationship between the inputs and the intended outputs. The network changes the weights of the connections during the training phase to minimize the discrepancy between the expected and real output. This process is continued until the network's accuracy reaches the required level. The capacity of ANNs to learn complex nonlinear correlations between inputs and outputs is one of its main advantages. This makes them especially effective in cases where the connection between inputs and outputs is unclear or cannot be represented using typical statistical approaches [57]. ANNs have been used, for example, to categorize pictures, identify speech, and estimate the likelihood of an event occurring based on a set of inputs [58]. The capacity of ANNs to generalize new data is another benefit. Once trained on a collection of examples, an ANN may be used to generate predictions on fresh, previously unknown data. The advantage of using ANNs for RBP binding site prediction lies in their ability to learn correlations and complex patterns in data without the need for explicit feature engineering. This means that ANNs can be trained on biological sequences and are not limited to predefined features set by the researcher. Moreover, ANNs demonstrate good generalization to new data, allowing them to predict binding sites for RBPs with limited experimental data [13].

To predict RNA binding protein binding locations, various deep learning methods utilizing ANNs have been developed. Convolutional neural networks (CNNs) have proven to be effective in identifying sequence motifs indicative of RBP binding preferences. CNN-based models, such as DeepBind [59], DeepSEA [60], and Basset [61], have been widely applied to genomic data for RBP binding site prediction. Another approach is the use of recurrent neural networks (RNNs) in RBP binding site prediction. RNNs are designed for handling sequential data, such as RNA sequences: models like deepTarget [62], deepMiRGene [63], and DeepNano [64] leverage RNNs to capture the temporal relationships between nucleotides and predict the likelihood of binding sites. Hybrid models that combine CNNs and RNNs have also been developed for RBP binding site prediction. These hybrid models exploit the strengths of both CNNs and RNNs to learn sequence motifs as well as capture the temporal relationships required for accurate RBP binding prediction, of which SPEID [65] is an example that demonstrates their highly competitive performance.

In summary, artificial neural networks, including CNNs, RNNs, and hybrid models, have become powerful tools in predicting RNA binding protein binding locations. They can learn complex patterns, generalize to new data, and do not rely on explicit feature engineering. These capabilities make ANNs invaluable for uncovering the sequence and structural properties indicative of RBP binding preferences.

3. Materials and Methods

The programming was performed using Jupyter Notebooks, which are publicly available at https://github.com/sofiamrtll/thesis. The organisation of it is available in the README file. The folders have been organised following the suggestions of 'The Good Research Code Handbook'[66].

3.1. Datasets

3.1.1. Sun et al.(RNA-seq)

[67] The RNA-seq study by Sun et al. was conducted using HEK293 cells and encompassed a total of 57,905 transcripts, with 32,396 transcripts exhibiting expression. The dataset was initially aligned to the hg19/GRCh37 genome version. However, to align it with other relevant files in the more recent hg38/GRCh38 genome version, a gene-to-gene mapping step was carried out (as described in the R script, available on the GitHub Repository). The CSV file, named GSE122425all.counts.293vs293NK.edgeRall, provides comprehensive details including gene IDs, lengths, HGNC symbols, and read counts from three distinct sequencing runs (SEQ1, SEQ2, SEQ3). Both absolute read counts and RPKM values (Reads Per Kilobase Million) are provided, but for the specific aim of this investigation, only the absolute read counts are considered relevant. This is because the absolute read count provides a more direct and informative measure for discriminating between expressed and non-expressed genes.

3.1.2. ENCODE(RNA-seq)

[68] The ENCODE RNA-seq study focused on HepG2 cells, encompassing a dataset of 207,507 transcripts, with 96,044 transcripts exhibiting expression. The accompanying CSV file contains extensive information including gene ID, length, transcript ID, effective length, and expected count. Additionally, the file provides values for TPM (Transcripts Per Kilobase Million) and FPKM (Fragments Per Kilobase Million), along with statistical data such as standard deviation and coefficient of quartile variation, which are not pertinent to the scope of this particular study.

3.1.3. ENCODE(eCLIP)

[69] The ENCODE eCLIP study encompassed HepG2 and HEK562 cells, investigating 223 RNA-binding proteins (RBPs) and their corresponding binding sites within the transcriptome, 73 of them were shared by both cell lines. Each RBP dataset consists of approximately 20,000

binding site coordinates in .bed format, specifying the RNA strand's chromosome, start nucleotide and end nucleotide. The files also provide supplementary details, including the gene coordinates associated with each binding site and various gene identifiers.

3.1.4. Linder et al.(miCLIP)

[31] The miCLIP study conducted by Linder et al. focused on HEK293 cells, providing 9278 coordinates representing m6A sites within the transcriptome. The data initially utilized the hg19/GRCh37 genome version, necessitating a subsequent genome liftover procedure to hg38/GRCh38 using UCSC LiftOver(http://genome.ucsc.edu)[70]. The dataset is in .bed format and includes the precise genomic coordinates of the m6A sites on the RNA strands, denoted by the chromosome, start nucleotide, and end nucleotide. Supplementary information such as the gene coordinates and various gene identifiers associated with each binding site is also provided.

3.2. Positive and Negative sampling for the RBP binding task

Positive events in CLIP-seq experiments correspond to observed binding, while negative events are defined by the absence of observed binding. Negative events are typically generated by uniformly sampling regions of the transcriptome that do not overlap with observed binding sites. This assumes that the likelihood of identifying cross-linked positions is equally probable across all transcriptome positions in the absence of RBP-specific binding features. The sampling approach used in this study follows the methodology described in the paper "A Systematic Benchmark of Machine Learning Methods for Protein-RNA Interaction" [13]. Two strategies are employed to generate negative samples for training and evaluation. The first strategy, referred to as negative-1, involves sampling positions from transcripts that overlap with at least one binding site of the protein of interest. The second strategy, known as negative-2, involves sampling from binding sites of other RBPs that were experimentally assessed in the dataset. This ensures that CLIP-seq biases are equally present in both the positive and negative sets, eliminating the informative nature of these biases in distinguishing positive and negative classes. By using negative-2, the models are prevented from learning these biases during training. Negatives were generated at a 1:1 ratio relative to the number of positive instances for each experiment. All methods were trained and evaluated separately using both sets of negatives.

3.3. Data Preprocessing

The current study operates under the assumption that RNA-binding protein (RBP) binding sites and N6-methyladenosine (m6A) sites exhibit a degree of conservation across different cell lines [10, 71, 72]. To begin, a set of expressed genes (number of reads > 0) common to the two investigated cell lines was generated, comprising a total of 20,954 transcripts. This step is essential as the CLIP files correspond to different cell lines, necessitating the exclusion of incompatible genes or transcripts from both datasets to facilitate the utilisation of cross-line data. Following this initial step, the miCLIP dataset was reduced to 9,087 features, while the eCLIP dataset underwent an average reduction of 1% for each investigated RBP. These steps are available in the Jupyter Notebook 'Preprocessing'. The resulting dataset for each RBP constitutes the positive set, which was subsequently divided equally into five folds for cross-validation purposes, see Figure 3.3.

The binding site coordinates were then extended to a fixed length of 400 nucleotides based on literature suggesting that longer sequences facilitate the prediction of binding sites [13]. Additionally, the previously sampled negative sets, namely 'negative-1' and 'negative-2,' were added to the corresponding RBP folders as part of the positive and negative sampling process. The files, initially in .bed format, were transformed into .fasta format for subsequent one-hot encoding of the sequences using the BedTools package (Figure 3.1). With the same package, the file containing the extended sequences was compared to the preprocessed miCLIP dataset in order to identify the N6-methyladenosine (m6A) sites present in each file. The entirety of the commands used is to be found in the Jupyter Notebook 'Encoding'.

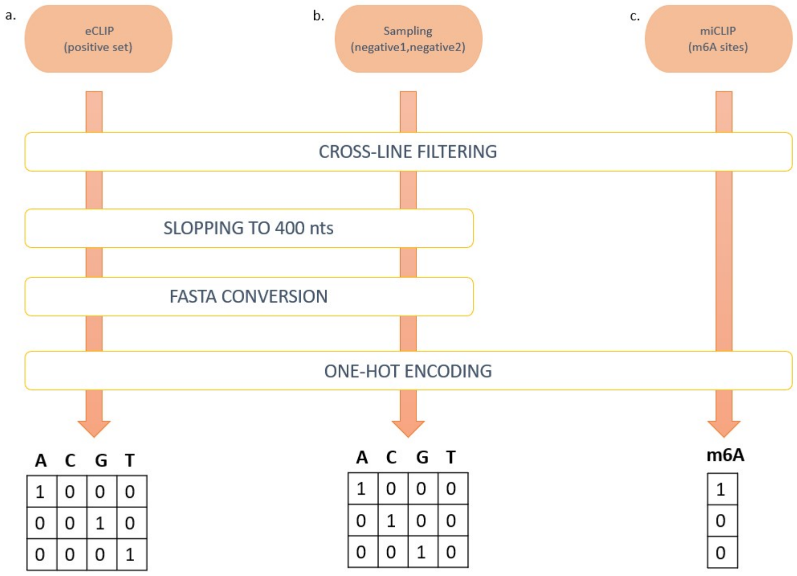


Figure 3.1.: Schematic Representation of the Preprocessing step.

a. eCLIP dataset [69] b. Negative sampling [13]c. miCLIP dataset [31].

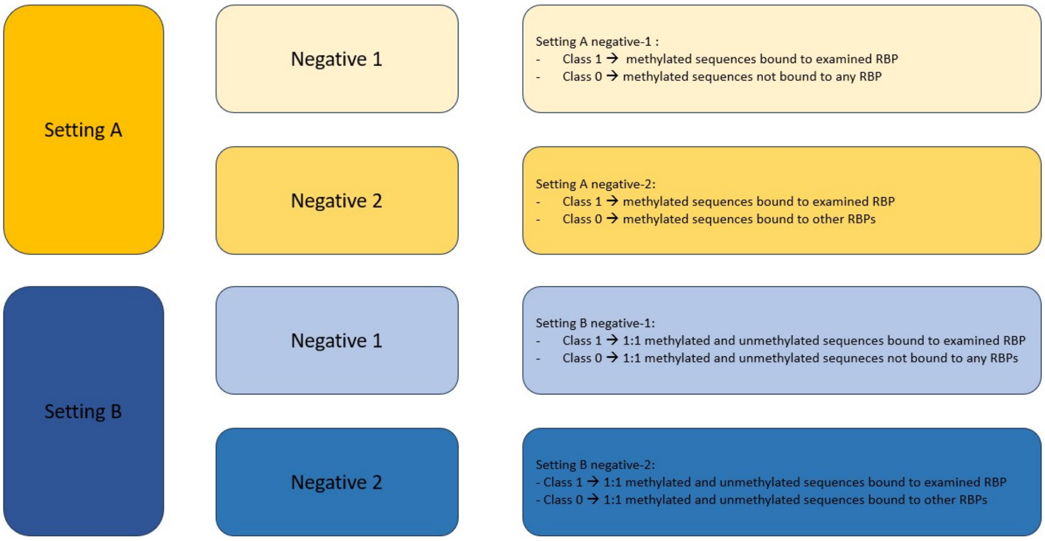
To address the issue of insufficient m6A site-containing dataset size, particularly in the negative1 class, the eCLIP dataset underwent augmentation by introducing a random number of nucleotides on both sides of the binding site, standardizing the sequences to a length of 400 nucleotides, differently from the method utilized to produce the sequences previously, which maintained the RBP binding site in the centre of the sequence. A more precise augmentation method was later employed, wherein the central nucleotide between the binding site and an m6A site served as the starting position for the expansion of the sequence, i.e. addition of a random number of nucleotides on both sides of the identified point. This ensured that the majority of augmented sequences contained both binding and m6A sites. Both augmentation approaches are available in the Jupyter Notebook ’Encoding’.

The augmented sequences were then converted to .fasta format, and the m6A sites were identified using the aforementioned methods. Subsequently, the .fasta sequences were encoded using one-hot encoding, representing the nucleotides A, C, G, and T as binary vectors. For the m6A integrated data, an additional vector was appended, containing the encoding for the m6A sites. Sequences of the same length were stacked, resulting in three-dimensional arrays specific to each class. These sequences were then labelled accordingly, and the different classes were combined and shuffled while maintaining the integrity of the sequences. These steps are programmed in the file model.py, in particular in the function ’prepare_raw_dataset()’.

3.4. Positives and Negatives for the Integrated Model

For the preparation of the datasets used as input for the model, two distinct settings were devised, each representing subsets derived from the original dataset consisting of positive samples, negative-1 samples, and negative-2 samples. A visualized in Figure 3.2, Setting A focuses on the subset comprising exclusively methylated sequences. Within Setting A, the ’Setting A - negative-1’ subset is composed of negative-1 samples, while the ’Setting A - negative-2’ subset leverages the negative-2 class that contains m6A-containing sequences as the negative class. It is important to note that after the introduction of augmented data, all these subsets also include the augmented samples, and they are denoted by adding ’-aug’ at the end of the subset name.

On the other hand, Setting B represents a subset that maintains a balanced 1:1 ratio between sequences containing m6A sites and sequences without m6A sites. Similar to Setting A, ’Setting B - negative-1’ employs the filtered negative-1 class, and ’Setting B - negative-2’ utilizes a subset of the negative-2 class. As with Setting A, these subsets are also extended to incorporate augmented data and are indicated by the addition of ’-aug’ in the subset name. The code used to achieve these different Settings is contained in the file model.py, in particular in the functions ’setting_A()’ and ’setting_B()’.



Setting A negative-1 :
- Class 1 → methylated sequences bound to examined RBP
- Class 0 → methylated sequences not bound to any RBP

Setting A negative-2:
- Class 1 → methylated sequences bound to examined RBP
- Class 0 → methylated sequences bound to other RBPs

Setting B negative-1:
- Class 1 → 1:1 methylated and unmethylated sequences bound to examined RBP
- Class 0 → 1:1 methylated and unmethylated sequneces not bound to any RBPs

Setting B negative-2:
- Class 1 → 1:1 methylated and unmethylated sequences bound to examined RBP
- Class 0 → 1:1 methylated and unmethylated sequences bound to other RBPs

Figure 3.2.: Schematic Representation of the different Settings.

3.5. Model Architecture

The model design was performed using the Keras package [73] and consists of a sequential Convolutional Neural Network (CNN), which is a widely used architecture for analysing sequential data. In particular, the untuned architecture is taken from the pysster [74] package, which is a Python package optimized for the training and interpretation of convolutional neural networks on biological sequence data.

The CNN comprises an initial Input layer, followed by a Convolutional layer, a Maxpooling layer, and a Dropout layer. This combination is repeated twice to enhance the model's ability to recognize essential features in the data. Subsequently, a Flatten layer is employed to convert the output into a vector representation. To exclude superficial information and focus on relevant features, a configuration involving double Dropout and Dense layers is utilized. The Dropout layers randomly deactivate neurons during training, while the Dense layers densely connect neurons to capture complex patterns in the data. This configuration aids in excluding non-essential information from the model. The final Dense layer ensures that the model's output is suitable for a binary classifier, providing a unique answer. This is achieved by utilising a single neuron in the output layer, which is typical for binary classification tasks. The architecture of the Model, tuned and untuned is available in Figure A.2, while the coding of the model architecture is to be found in the file model.py.

3.6. Model Training

Initially, the model employs default parameters obtained from the pysster library [74]. However, to optimize performance for each RNA-binding protein (RBP), the architecture and hyperparameters are subsequently tuned. This tuning process is carried out for each RBP individually to develop a specific model that accurately recognizes the unique binding sequence associated with each RBP. The architecture optimization and hyperparameter tuning have been performed using a RandomSearch tuner with search space defined as in Figure A.1. The hyperparameters have been carefully tuned to achieve the highest validation accuracy. Detailed information can be found in the 'model.py' file, specifically in the 'prepare_best_model()' function.

In order to mitigate overfitting, an Early Stopping mechanism has been implemented. Additionally, class weights have been incorporated to address any class imbalance issues. Two optimized models were saved for the analysis, each designed to handle specific input shapes. The first model has an input shape of (400, 4), corresponding to pure sequence datasets, while the second model has an input shape of (400, 5), accommodating m6A integrated datasets. Both models share the same optimized hyperparameters, ensuring comparability in their performance evaluations. Following the analysis of the obtained results, a specific set of hyperparameters was arbitrarily selected from the set of most occurring hyperparameters across all tuned models, they can be found in Table A.2. The choice of the optimizer was Adam optimizer, which is widely used; its parameters are to be found in Table A.1, while the loss function employed is Binary CrossEntropy, due to its suitability for binary classification tasks.

These hyperparameters were then employed consistently across all subsequent runs without differentiation between proteins. This approach aimed to maintain uniformity and fairness in the evaluation of the model's performance across different protein datasets. The model is to be found in the file model.py in the function 'create_baseline_model()'. For all datasets, the splitting process was carried out using an 80-20 split ratio to create separate training and testing datasets. Furthermore, within the training set, an additional 80-20 split was performed to create training and validation subsets. The schematic representation of this dataset splitting procedure can be observed in Figure 3.3.

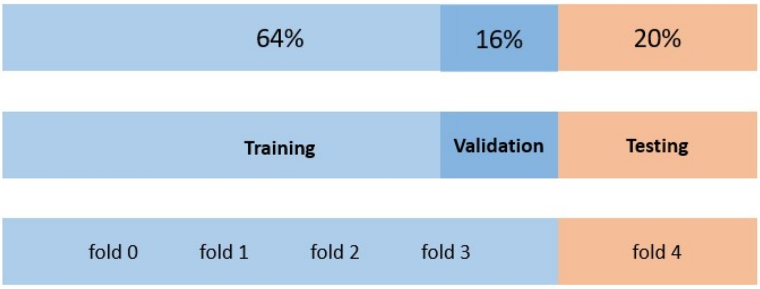


Figure 3.3.: Schematic Representation of the Dataset Split.

4. Results

The plots and tables used in this manuscript are produced in the Jupyter Notebook 'Plots' and 'Aggregated_Evaluation': all of the code is available there.

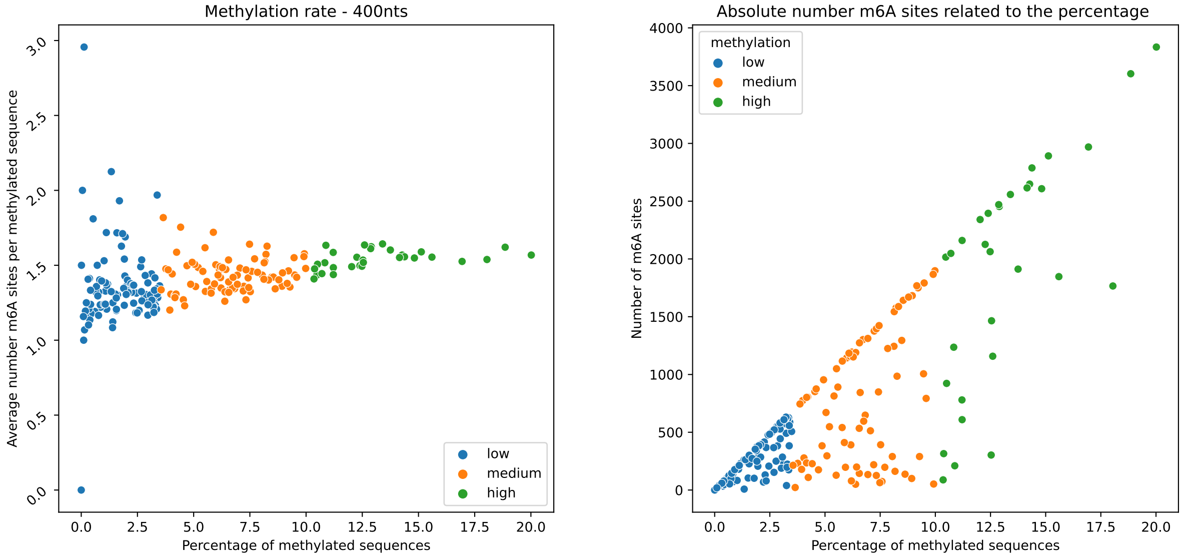
4.1. Preliminary Analysis

An initial analysis of the dataset composition was conducted to gain insights into its characteristics. This involved comparing the original number of sequences to the number of sites that belong to expressed genes in the considered cell line, which refers to the filtered dataset used in this study, across each set of RNA-binding proteins (RBPs) (see Figure 4.2). The percentage of sequences containing non-expressed genes was found to be consistent across different RBP sets, with an average of approximately 1%. These sequences were filtered out to facilitate the utilization of cross-line data, as the CLIP files correspond to different cell lines, the exclusion of incompatible genes or transcripts from both datasets is necessary.

Subsequently, the comparison was made between the total number of sequences (i.e., expressed sequences) and the number of methylated sequences, which are sequences containing at least one N6-methyladenosine (m6A) site (see Figure 4.3). The subset of methylated sequences was found to be relatively small, with an average percentage of methylated sequences of 5.16% on the total dataset. The maximum percentage of m6A occurrence was observed in PABPC4 (K562 cell line) at 20%.

The absolute number of m6A sites within each RBP set was further examined (see Figure 4.5). It was revealed that m6A sites are infrequent, with the highest number of m6A sites found in PABPC4 (K562 cell line) at 3834, while SUGP2 (HepG2 cell line) did not contain any m6A sites. The average number of m6A sites across all RBP sets was 675.56.

The average number of m6A sites per sequence was investigated both across all sequences (see Figure 4.7) and specifically among methylated sequences (see Figure 4.6). The maximum value was observed in PABPC4 (K562 cell line) at 3.13, while the mean value of this metric was 0.07 m6A sites per sequence. When considering only the methylated sequences, the mean value was 1.4, with the maximum value of 2.9 in EXOSC5 (K562 cell line).

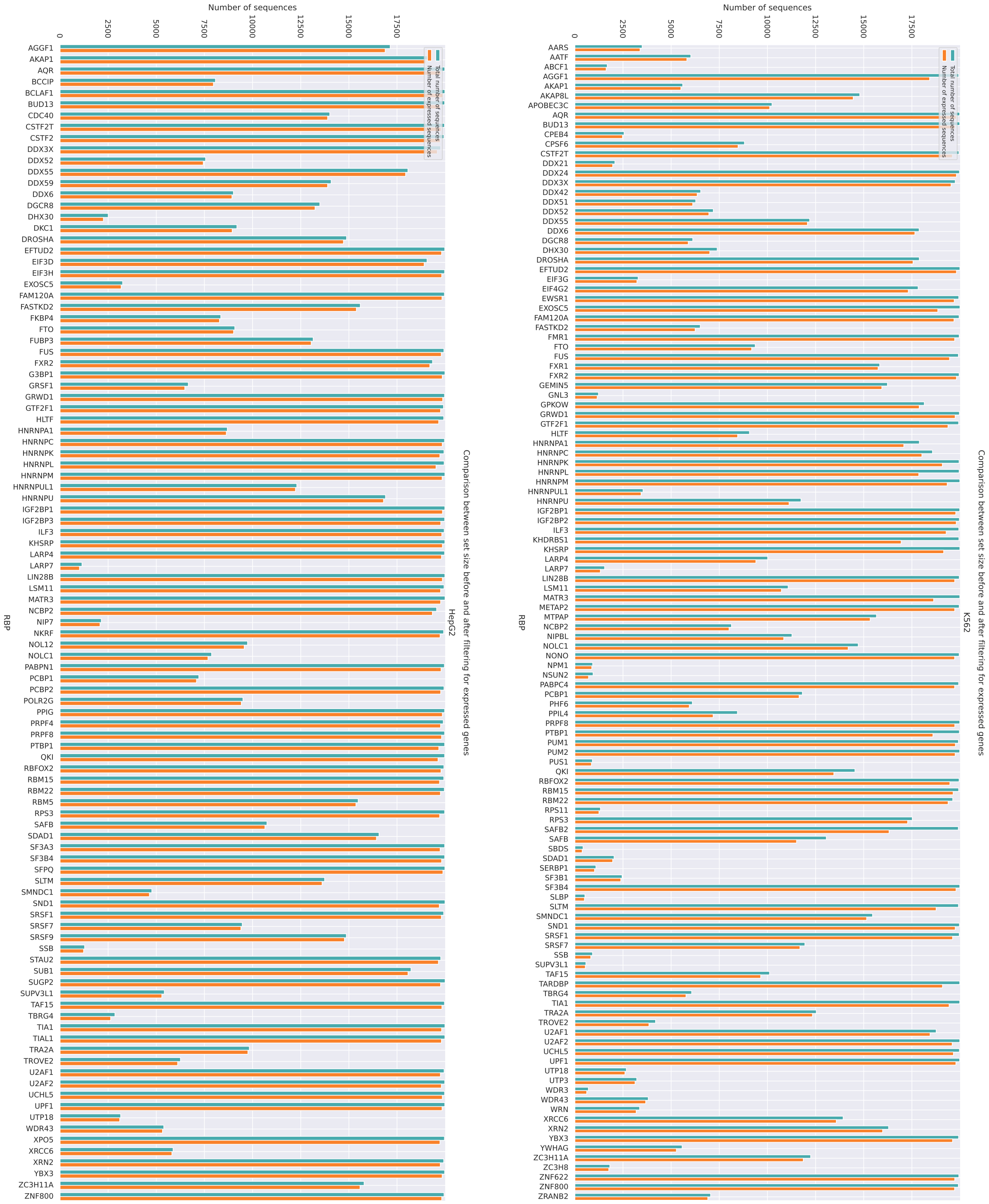


(a) (b)

Figure 4.1.: Distribution of Methylation Rates related to the Number of m6A Sites

Additionally, two scatterplots in Figure 4.1 illustrated the distribution of methylation rates using different metrics. In 4.1a, the methylation rate was related to the average number of m6A sites per methylated sequence, revealing that lower methylation rates were associated with a higher number of m6A sites per methylated sequence. High methylation rate protein sets showed a more concentrated distribution around the average of 1.4 m6A sites per methylated sequence. In 4.1b, the methylation rate was related to the absolute number of m6A sites present in the protein sets, highlighting that some high methylation protein sets had a smaller number of m6A sites compared to the medium and low methylation rate protein sets. These analyses encompassed all RBP sets in the dataset, providing a comprehensive overview of the chosen dataset. It is important to note that these investigations were conducted solely on the positive dataset containing RBP binding sites, and the negative datasets were not examined.

Following the introductory investigation, not all of the 223 RBPs in the dataset were processed. A subset of proteins with a high methylation rate, specifically those with more than 10.5% of methylated sequences relative to the total number of sequences, were identified. Additionally, proteins whose input sets resulted void after filtering for m6A sites were removed, namely SLBP (K562 cell line), SERBP1 (K562 cell line), SBDS (K562 cell line), and RPS11 (K562 cell line). In the subsequent sections of this chapter, the analysis focuses on this selected set of 31 proteins (see Table A.3), which were processed and used as input for the model.

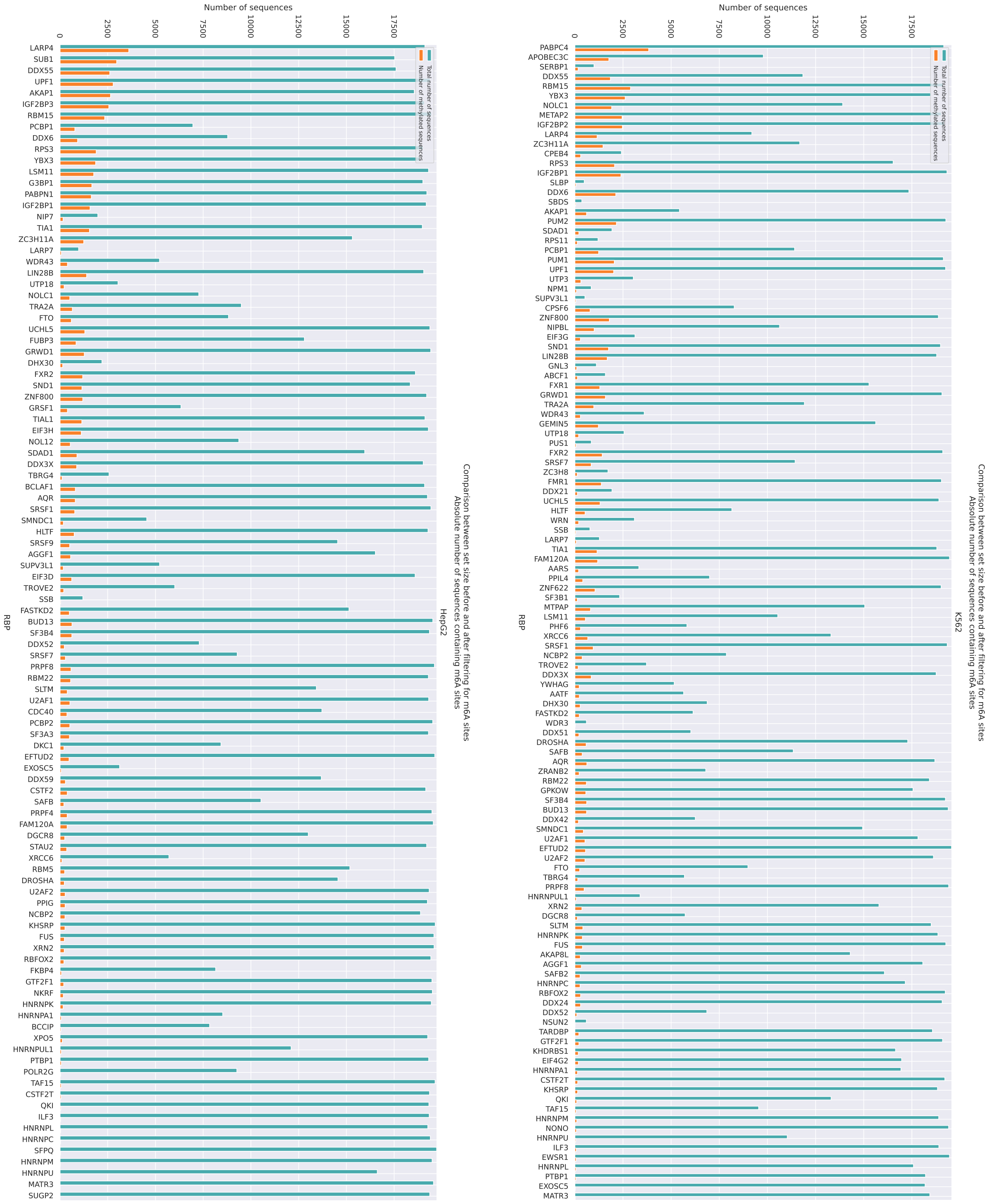


(a) HepG2 cell line

(b) K562 cell line

Figure 4.2.: Comparison of Total and Expressed Sequences

Comparative analysis of total sequences (blue) and expressed sequences (orange) in each protein set (RBP), sorted alphabetically.

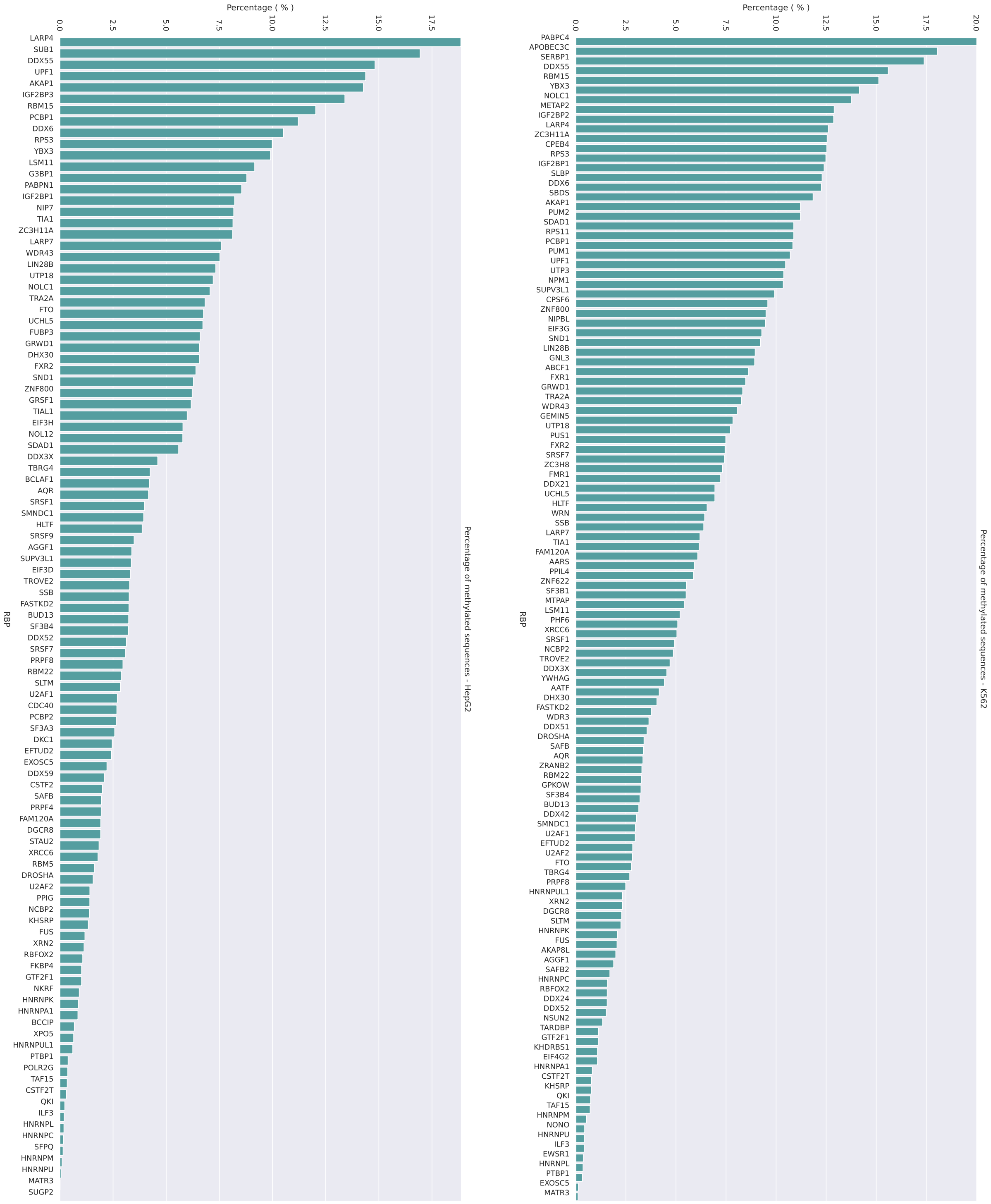


(a) HepG2 cell line

(b) K562 cell line

Figure 4.3.: Comparison of Total and Methylated Sequences

Comparative analysis of total sequences (blue) and methylated sequences (orange) in each protein set (RBP), sorted by descending methylation rate.

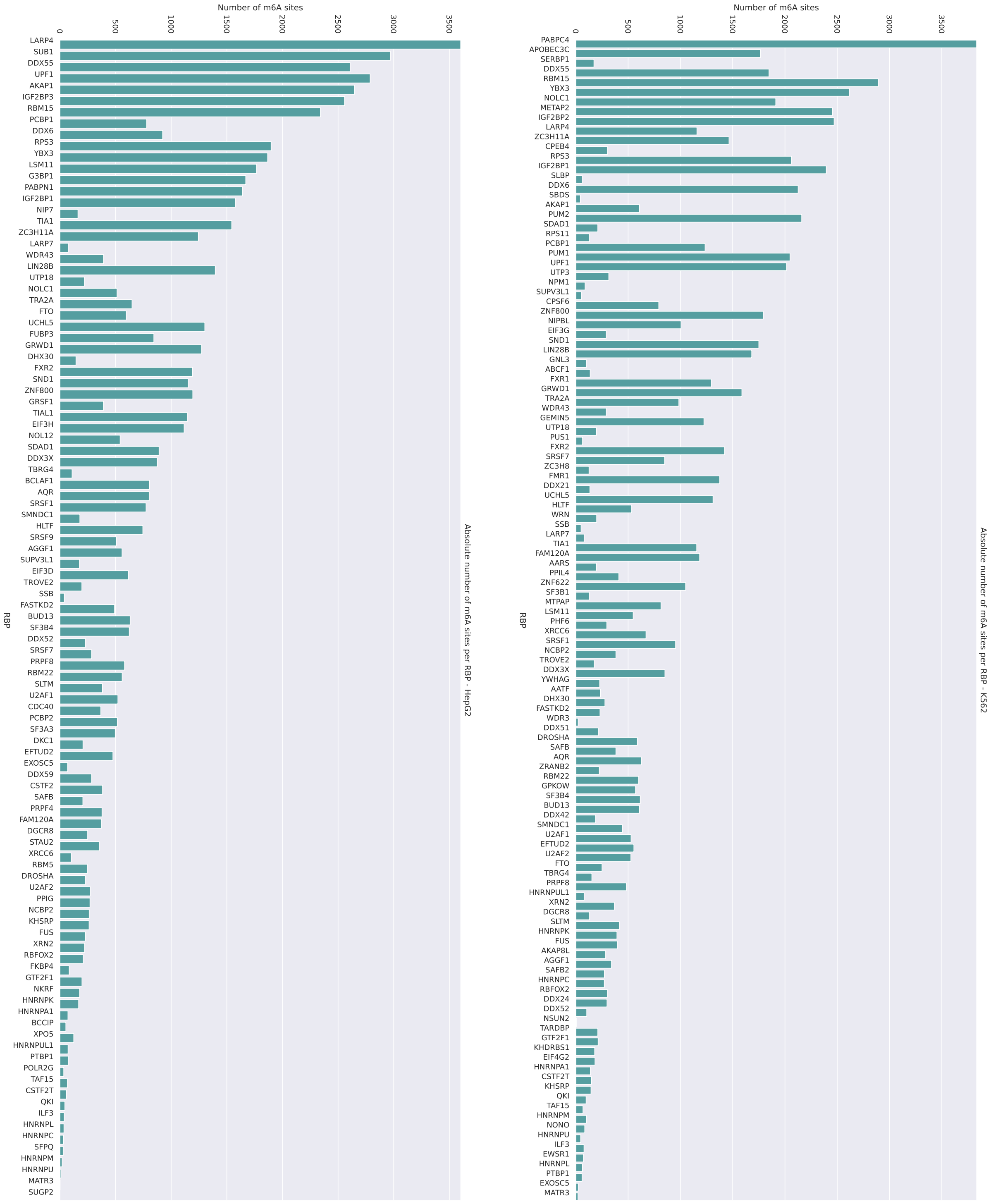


(a) HepG2 cell line

(b) K562 cell line

Figure 4.4.: Ranked Distribution of Methylated Sequences in Protein Sets.

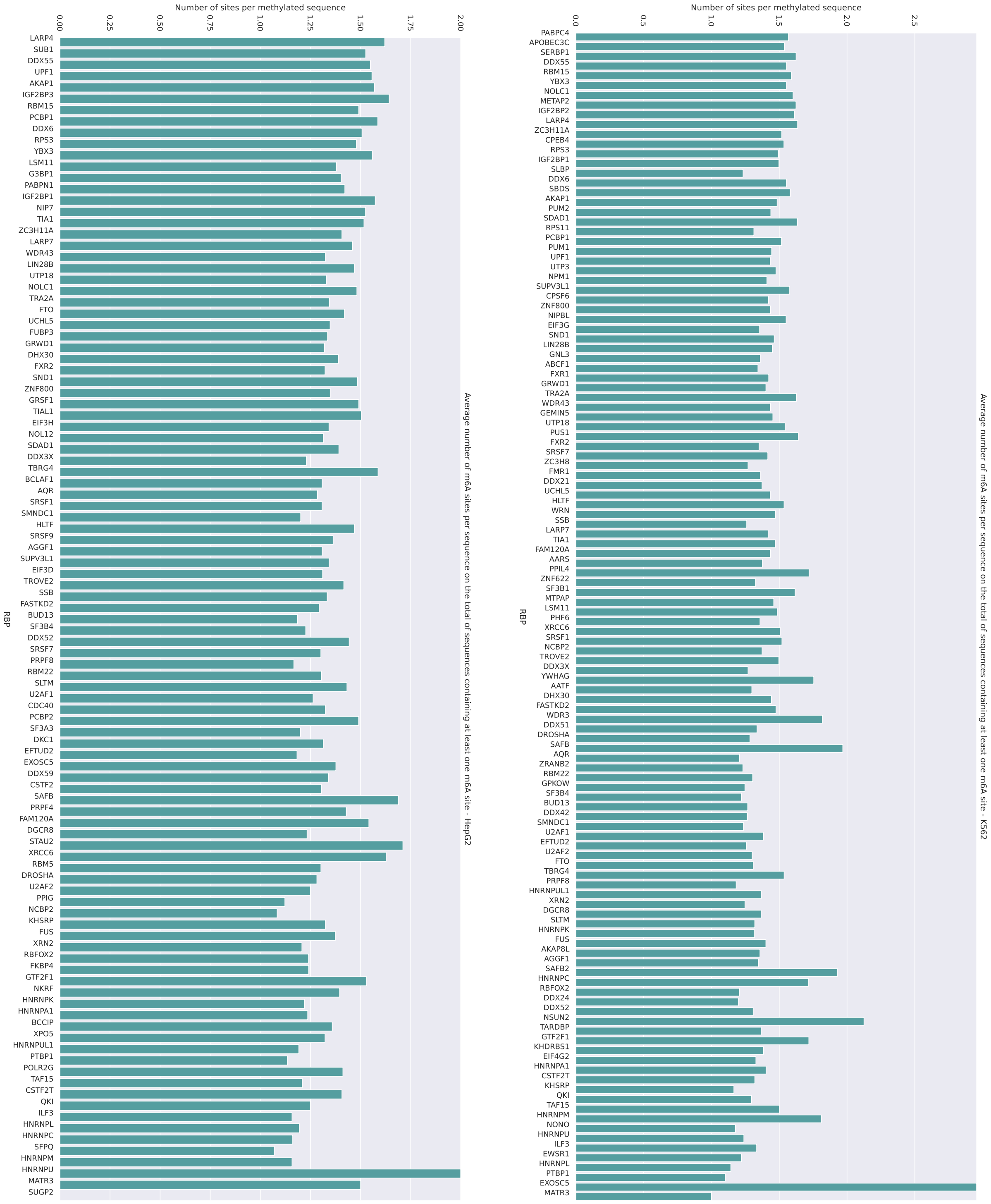
Methylated sequence percentages in protein sets, ranked by methylation rates, offer a comprehensive view of relative abundance. The ranking identifies protein sets with varying methylation levels, highlighting variability in methylation patterns across samples.



(a) HepG2 cell line

(b) K562 cell line

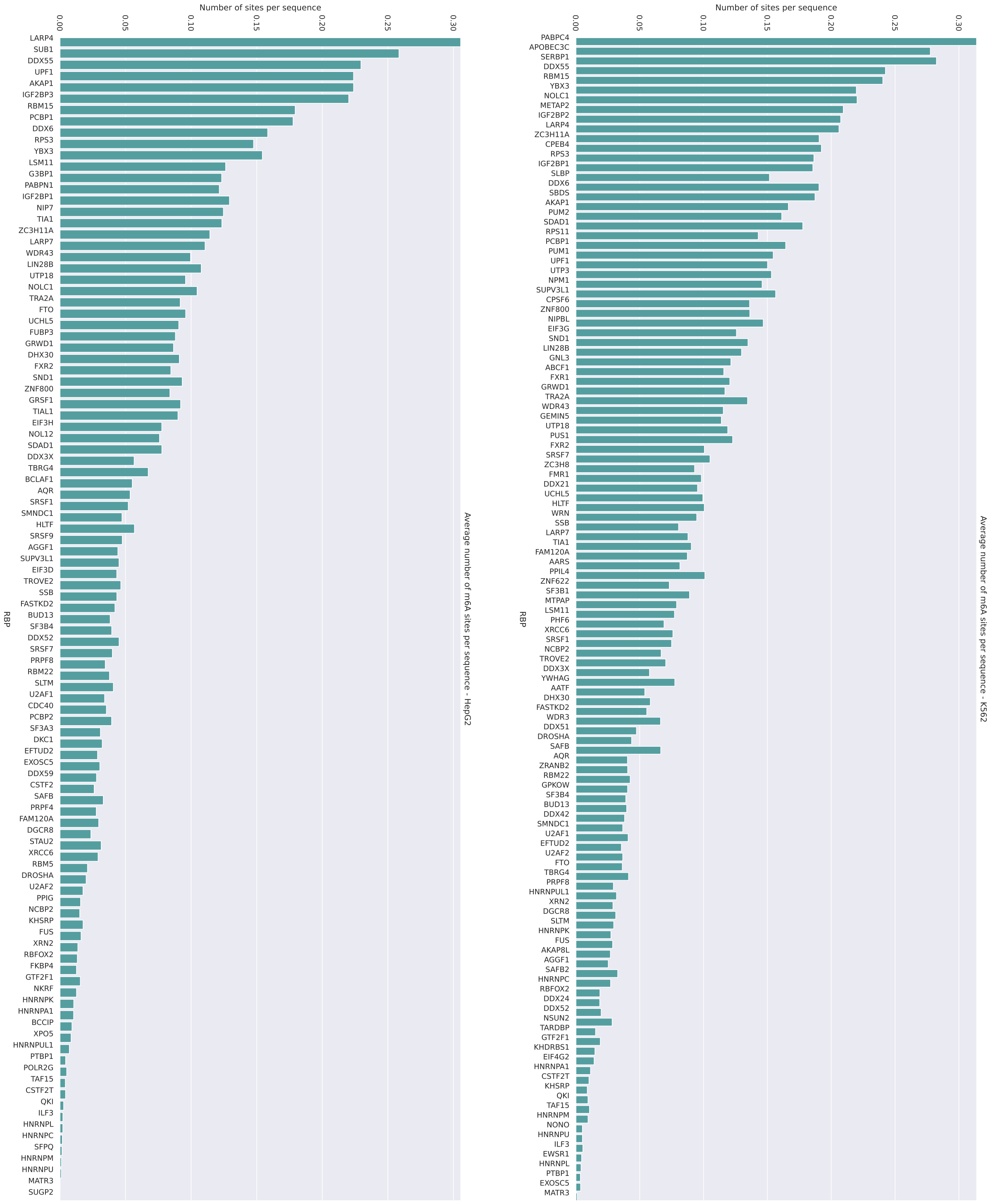
Figure 4.5.: Quantification of m6A Sites in Each RBP Set: Ranking Based on Methylation Rates



(a) HepG2 cell line

(b) K562 cell line

Figure 4.6.: Average Number of m6A Sites per Sequence among Methylated Sequences, Ranked by Methylation Rates



(a) HepG2 cell line

(b) K562 cell line

Figure 4.7.: Average Number of m6A Sites per Sequence among all Sequences, Ranked by Methylation Rates

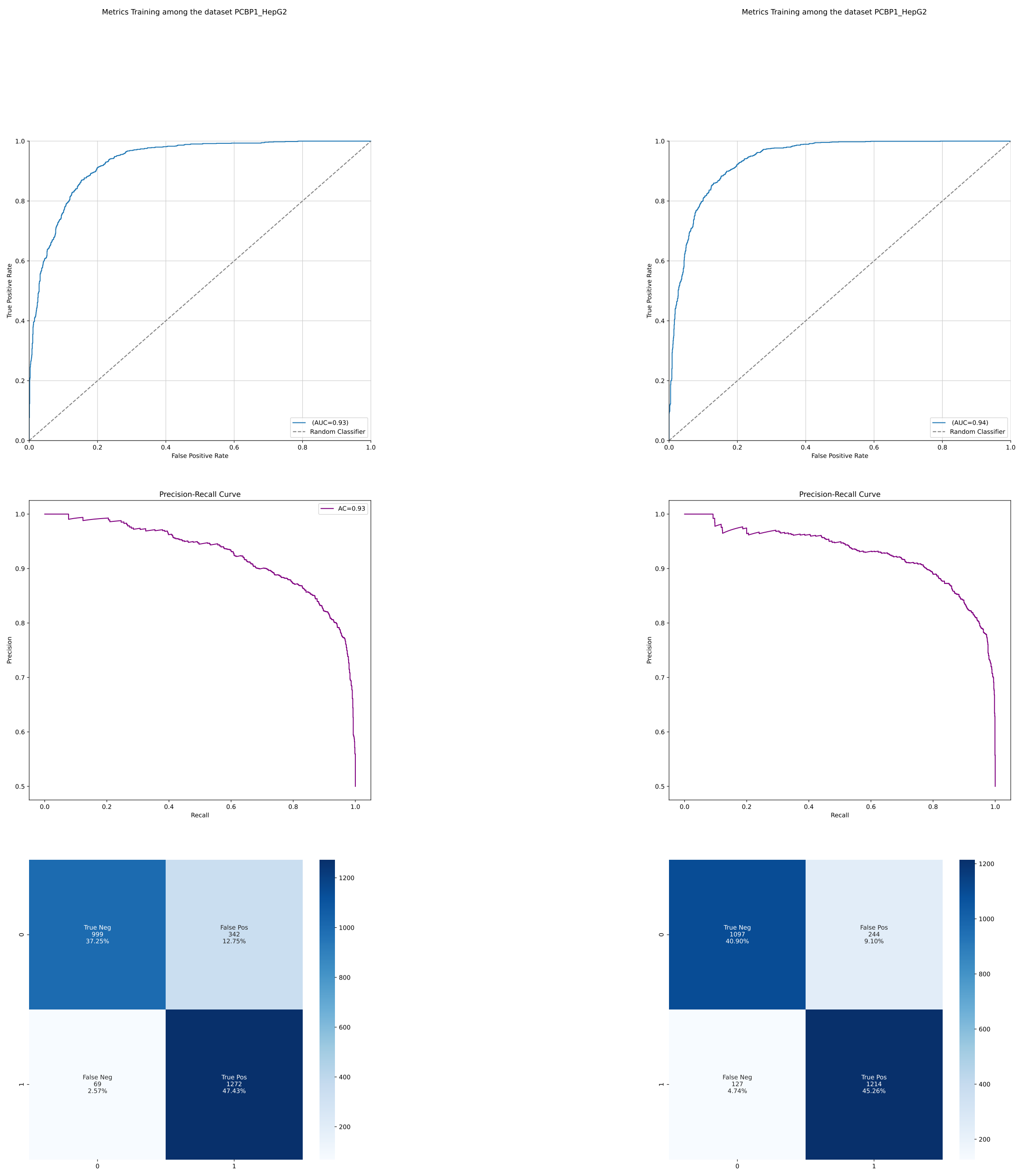
4.2. Tuning of the Model

In this section, a comparison is conducted between the standard model and the tuned model. The comparison is performed using the entire RNA-binding protein (RBP) dataset, specifically the expressed sequences, while excluding information related to methylations. The selected subset of 31 proteins (see Table A.3) is chosen based on their methylation rates, as they are anticipated to exhibit detectable influences of m6A methylation, thereby enabling clearer classification for the model. One protein, PCBP1, from the HepG2 cell line, is chosen as a representative example to showcase the results, with a methylation rate of 11.2% and an absolute number of 780 m6A sites in the positive set, which is close to the average across all protein sets.

The collected results include the accuracy, loss, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) of both models across training epochs, as presented in the Appendix A (Figure A.3). Furthermore, the analysis involves evaluating the area under the curve (AUC), precision-recall curve (PRC), and confusion matrix associated with the predictions, as shown in Figure 4.8. Notably, both models demonstrate satisfactory classification performance, with an AUROC and AUPRC of 0.93 for the untuned model and an AUROC of 0.94 for the tuned model. The tuned model achieves a correct classification rate of 86.16% for the sequences, while the untuned model achieves 84.68% accuracy. These results are specific to the protein set PCBP1 from the HepG2 cell line.

However, the aggregated results, depicted in entirety Table 4.1, provide a discussion on the average AUROC and AUPRC across all investigated proteins. The findings in Figure 4.9 align with the specific results: in fact, they display an average receiver operating characteristic (ROC) score of 0.86 and an average precision-recall (PR) AUC of 0.86 for the untuned model. For the tuned model, the mean values for PR AUC and ROC AUC are 0.85 and 0.86, respectively.

Regarding the analysis of sequences containing binding sites for a particular protein, both the standard model and the tuned model exhibit similar behaviours in terms of accuracy and loss throughout the training epochs, with comparable AUC and PRC values for the training set. However, in terms of prediction, the tuned model shows little to no optimization. Therefore, it was decided to artificially choose a set of hyperparameters from the tuned models, comparing the most frequently appearing ones. These will be used for all subsequent tests. This decision aims to enhance reproducibility and facilitate further research by other groups in this field.

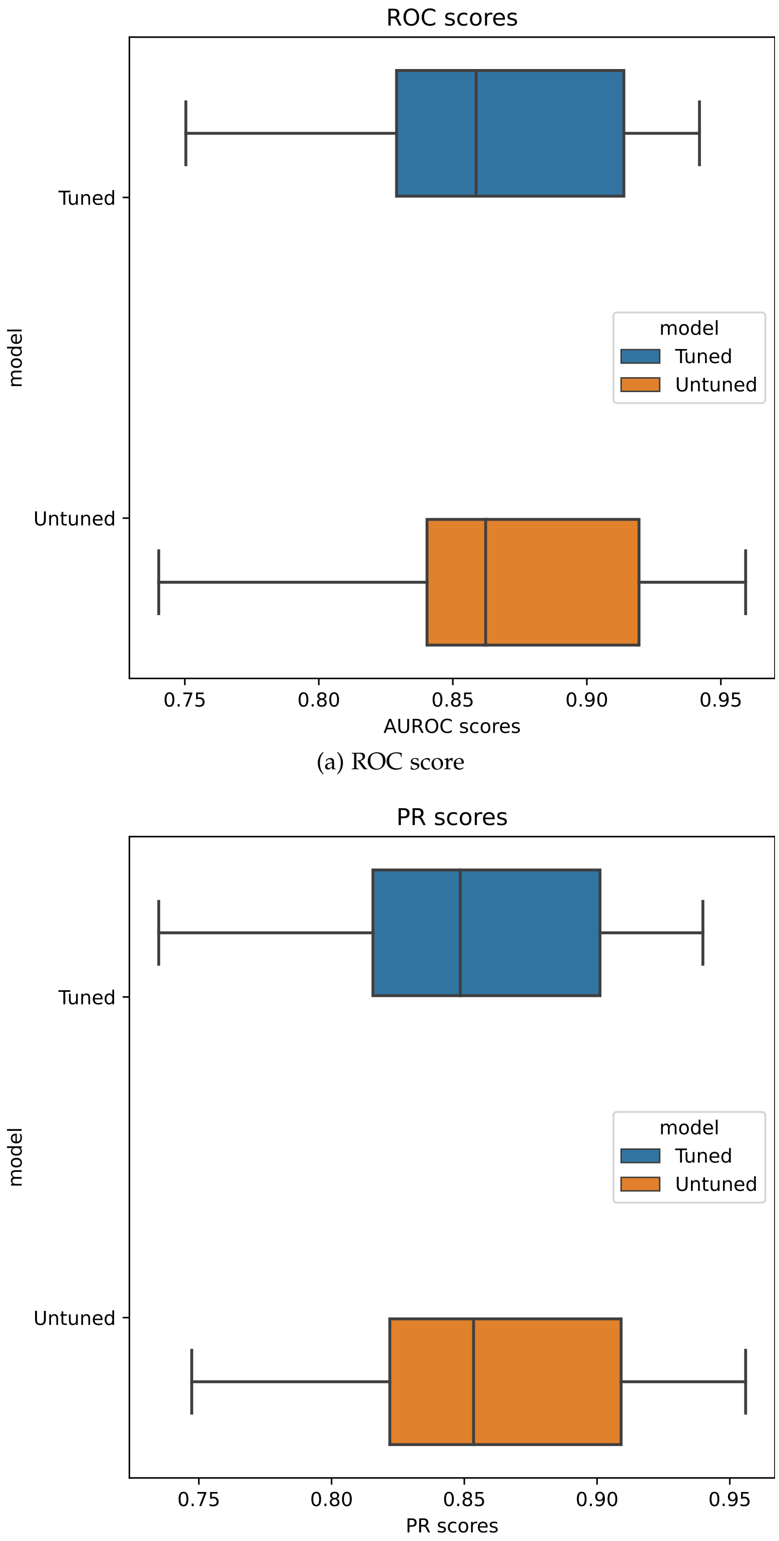


(a) Untuned Model

(b) Tuned Model

Figure 4.8.: Comparison of Untuned and Tuned Model: Prediction

Comparison of performance metrics for prediction between the untuned and tuned models: Area Under the Receiver Operating Characteristic Curve (AUROC) (top), Precision-Recall (PR) values (middle), and Confusion Matrix (bottom).



(a) ROC score

(b) PR score

Figure 4.9.: Comparison of Untuned and Tuned Model

Table 4.1.: Aggregated Evaluation for the Untuned and Tuned Model using the whole Dataset without Methylation Data Integration

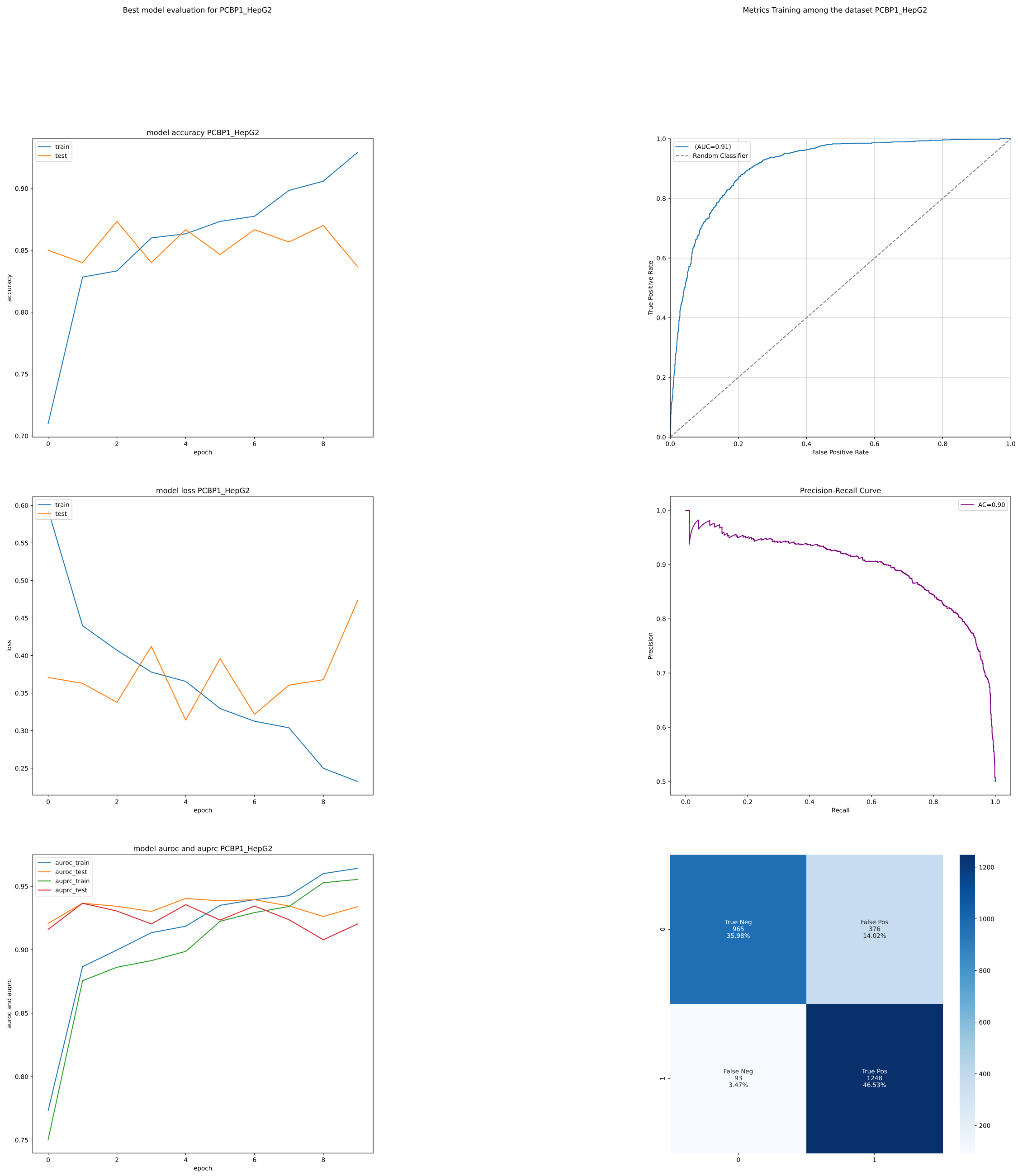
Table 4.2.: Untuned Model

Cell line	RBP	roc auc score	pr auc score
HepG2			
	AKAP1	0.8461	0.8323
	DDX55	0.8645	0.8480
	DDX6	0.8537	0.8285
	IGF2BP3	0.9180	0.9034
	LARP4	0.8565	0.8390
	PCBP1	0.9302	0.9189
	RBM15	0.9264	0.9170
	SUB1	0.8579	0.8384
	UPF1	0.8625	0.8476
K562			
	AKAP1	0.8134	0.8075
	APOBEC3C	0.8467	0.8458
	CPEB4	0.8083	0.8152
	DDX55	0.8624	0.8507
	DDX6	0.8697	0.8637
	IGF2BP1	0.8710	0.8617
	IGF2BP2	0.9229	0.9145
	LARP4	0.8515	0.8416
	METAP2	0.9248	0.9160
	NOLC1	0.8359	0.8185
	NPM1	0.7934	0.8084
	PABPC4	0.8535	0.8438
	PCBP1	0.9381	0.9299
	PUM1	0.9377	0.9354
	PUM2	0.9105	0.9117
	RBM15	0.9556	0.9508
	RPS11	0.8122	0.8088
	RPS3	0.9269	0.9154
	SBDS	0.7406	0.7473
	SDAD1	0.8988	0.9004
	SERBP1	0.8443	0.8576
	SLBP	0.7403	0.7708
	UPF1	0.8250	0.8195
	UTP3	0.8632	0.8426
	YBX3	0.8937	0.8890
	ZC3H11A	0.8020	0.7977

Table 4.3.: Tuned Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.8249	0.8097
	DDX55	0.8581	0.8423
	DDX6	0.8597	0.8402
	IGF2BP3	0.9126	0.8975
	LARP4	0.8741	0.8587
	PCBP1	0.9385	0.9278
	RBM15	0.9288	0.9154
	SUB1	0.8409	0.8066
	UPF1	0.8398	0.8277
K562			
	AKAP1	0.8025	0.7782
	APOBEC3C	0.8769	0.8725
	CPEB4	0.7535	0.7742
	DDX55	0.8567	0.8485
	DDX6	0.8726	0.8598
	IGF2BP1	0.8590	0.8447
	IGF2BP2	0.9278	0.9185
	LARP4	0.8375	0.8286
	METAP2	0.9150	0.9047
	NOLC1	0.8451	0.8228
	NPM1	0.8136	0.8149
	PABPC4	0.8215	0.8163
	PCBP1	0.9337	0.9256
	PUM1	0.9420	0.9399
	PUM2	0.9216	0.9237
	RBM15	0.9296	0.9235
	RPS11	0.8290	0.8089
	RPS3	0.9179	0.9097
	SBDS	0.7504	0.7349
	SDAD1	0.8885	0.8822
	SERBP1	0.8587	0.8731
	SLBP	0.7532	0.7954
	UPF1	0.8291	0.8239
	UTP3	0.8569	0.8486
	YBX3	0.8827	0.8810
	ZC3H11A	0.7993	0.7908

A supplementary test was carried out utilizing a downsized dataset for the protein PCBP1 (HepG2 cell line), with total datapoint count of 1500 for training and validation sets (80-20 split) and the tuned hyperparameters. The results align with the previous ones and can be seen in Figure 4.10.



(a) Training

(b) Predictions

Figure 4.10.: Training and Prediction Metrics for the downsized dataset Model

Comparison of performance metrics for prediction between the untuned and tuned models: Area Under the Receiver Operating Characteristic Curve (AUROC) (top), Precision-Recall (PR) values (middle), and Confusion Matrix (bottom).

4.3. Setting A

After conducting a thorough analysis of the label frequency in the dataset consisting exclusively of methylated sequences, it was observed that the dataset was very scarce, see Figure A.4. Recognizing the importance of a balanced and numerous dataset for effective training of a binary classifier, it was decided to exclude this dataset from the analysis. Therefore, the focus of this section is to present the results pertaining to Setting A with the augmented dataset. This decision ensures that the analysis is based on a dataset that adheres to the fundamental requirement of balance, enabling robust training and evaluation of the binary classifier.

The results presented in this section are organized into two subsections: *Negative1* and *Negative2*. Each subsection examines the performance of the integrated model, and for comparison purposes, a corresponding baseline model is also included. The purpose of the baseline model is to serve as a control, where both models are fed the same datasets, but the baseline model lacks the additional 5th channel that encodes the methylation data. This allows for a clear assessment of the impact of integrating methylation data on the model's performance. By comparing the performance of the integrated model to the baseline model, we can discern the influence of incorporating methylation data into the model. The results highlight any improvements or changes in performance that can be attributed to the inclusion of methylation information.

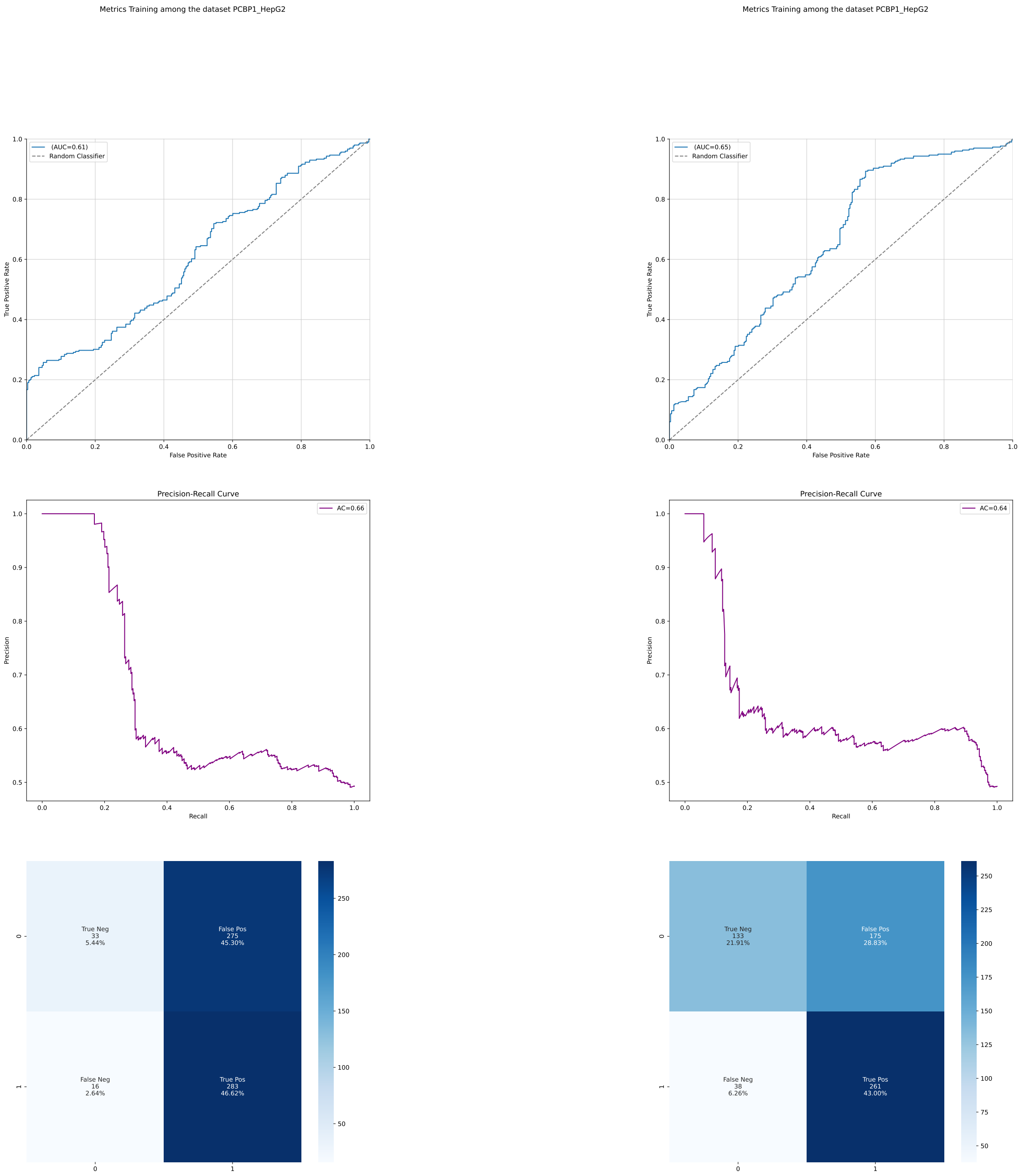
4.3.1. Negative 1

For the dataset consisting exclusively of methylated sequences and utilizing the negative-1 set as the negative class, the following results were obtained.

The collected results include the label frequency and numerosity of the dataset (Figure A.5), the accuracy, loss, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) for both the baseline and integrated models across the training epochs. These results can be found in the Appendix A (Figure A.7), and the observed curves suggest a potential case of overfitting, with loss values below 0.2 and accuracy approaching 1.0.

Furthermore, the analysis involves evaluating the area under the curve (AUC), precision-recall curve (PRC), and confusion matrix associated with the predictions, as depicted in Figure 4.11. It is notable that the classification performance decreased compared to the previous model that considered the entire dataset. The baseline model achieved an AUROC and AUPRC of 0.61 and 0.66, respectively, while the integrated model achieved an AUROC and AUPRC of 0.65 and 0.64, respectively. The baseline model achieved a correct classification rate of 52.06% for the sequences, while the integrated model achieved 64.91% accuracy. These results pertain specifically to the protein set PCBP1 from the HepG2 cell line.

However, the aggregated results, as presented in Table 4.4, provide a comprehensive discussion on the average AUROC and AUPRC across all investigated proteins. The findings depicted in Figure 4.12 align with the specific results, indicating an average receiver operating characteristic (ROC) score of 0.51 and an average precision-recall (PR) AUC of 0.66 for the baseline model. As for the integrated model, the mean values for PR AUC and ROC AUC are 0.52 and 0.53, respectively.

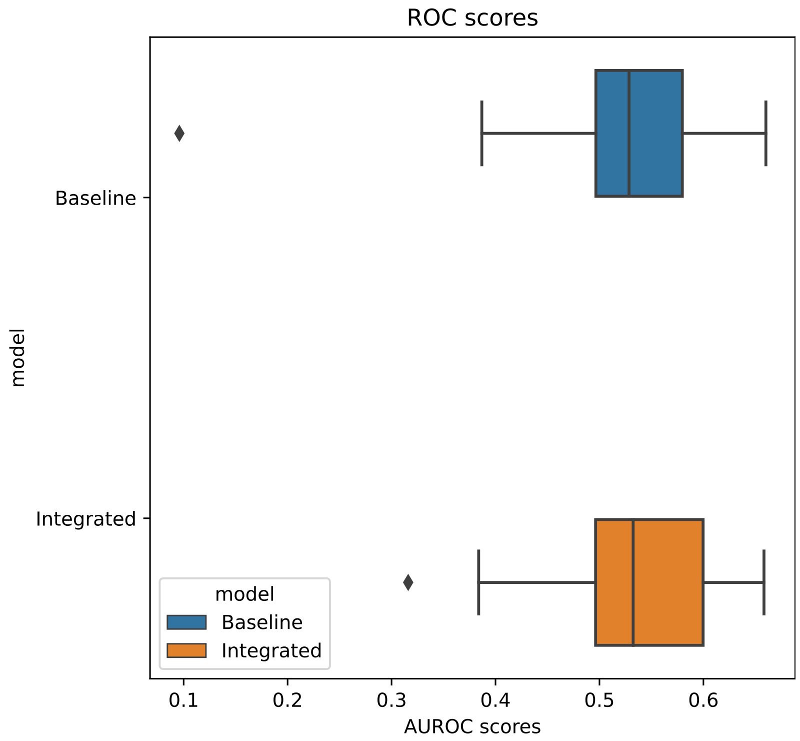


(a) Baseline Model

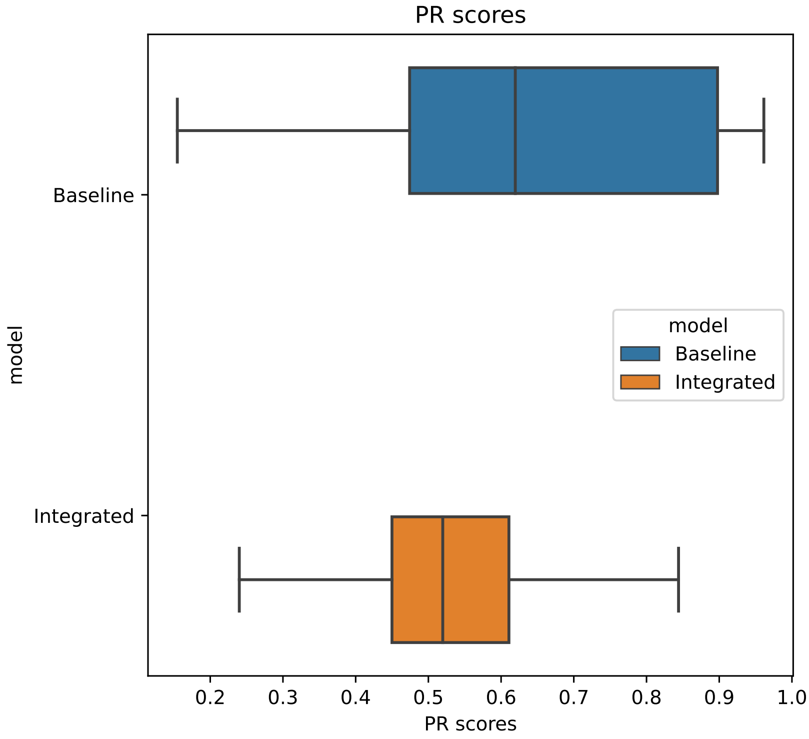
(b) Integrated Model

Figure 4.11.: Comparison of Baseline and Integrated Model: Prediction

Comparison of performance metrics for prediction between the baseline and integrated models: Area Under the Receiver Operating Characteristic Curve (AUROC) (top), Precision-Recall (PR) values (middle), and Confusion Matrix (bottom).



(a) ROC score



(b) PR score

Figure 4.12.: Comparison of Baseline and Integrated Model: Setting A negative-1

Table 4.4.: Aggregated Evaluation for Setting A - Negative1

Table 4.5.: Baseline Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.5813	0.6422
	DDX55	0.5000	0.4735
	DDX6	0.5463	0.4094
	IGF2BP3	0.5513	0.6123
	LARP4	0.5263	0.5312
	PCBP1	0.6083	0.6584
	RBM15	0.6162	0.5528
	SUB1	0.4088	0.4048
	UPF1	0.6601	0.7467
K562			
	AKAP1	0.6244	0.5794
	APOBEC3C	0.4175	0.4666
	CPEB4	0.6489	0.8830
	DDX55	0.4880	0.3749
	DDX6	0.4220	0.4495
	IGF2BP1	0.4678	0.5866
	IGF2BP2	0.4116	0.3633
	LARP4	0.5702	0.6121
	METAP2	0.5864	0.5866
	NOLC1	0.5639	0.6268
	NPM1	0.4619	0.4209
	PABPC4	0.4535	0.4765
	PCBP1	0.6230	0.5772
	PUM1	0.5453	0.4663
	PUM2	0.5030	0.5148
	RBM15	0.5606	0.5451
	RPS3	0.5030	0.3605
	SDAD1	0.3868	0.1547
	UPF1	0.5364	0.4588
	UTP3	0.4956	0.4465
	YBX3	0.5750	0.6579
	ZC3H11A	0.5851	0.6405

Table 4.6.: Integrated Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.6325	0.7010
	DDX55	0.4425	0.4407
	DDX6	0.4843	0.3634
	IGF2BP3	0.4458	0.5244
	LARP4	0.5498	0.5151
	PCBP1	0.6505	0.6379
	RBM15	0.5000	0.4928
	SUB1	0.5047	0.5059
	UPF1	0.5319	0.6185
K562			
	AKAP1	0.5974	0.5291
	APOBEC3C	0.3838	0.4370
	CPEB4	0.6126	0.8440
	DDX55	0.5823	0.4841
	DDX6	0.4954	0.4912
	IGF2BP1	0.4999	0.6376
	IGF2BP2	0.4966	0.4064
	LARP4	0.5329	0.6019
	METAP2	0.6349	0.6110
	NOLC1	0.4899	0.6107
	NPM1	0.3160	0.2400
	PABPC4	0.5127	0.4678
	PCBP1	0.6493	0.5413
	PUM1	0.5980	0.4871
	PUM2	0.5348	0.5294
	RBM15	0.6583	0.6009
	RPS3	0.5587	0.3965
	SDAD1	0.6048	0.2611
	UPF1	0.4821	0.4370
	UTP3	0.5142	0.4530
	YBX3	0.5424	0.6115
	ZC3H11A	0.5208	0.6093

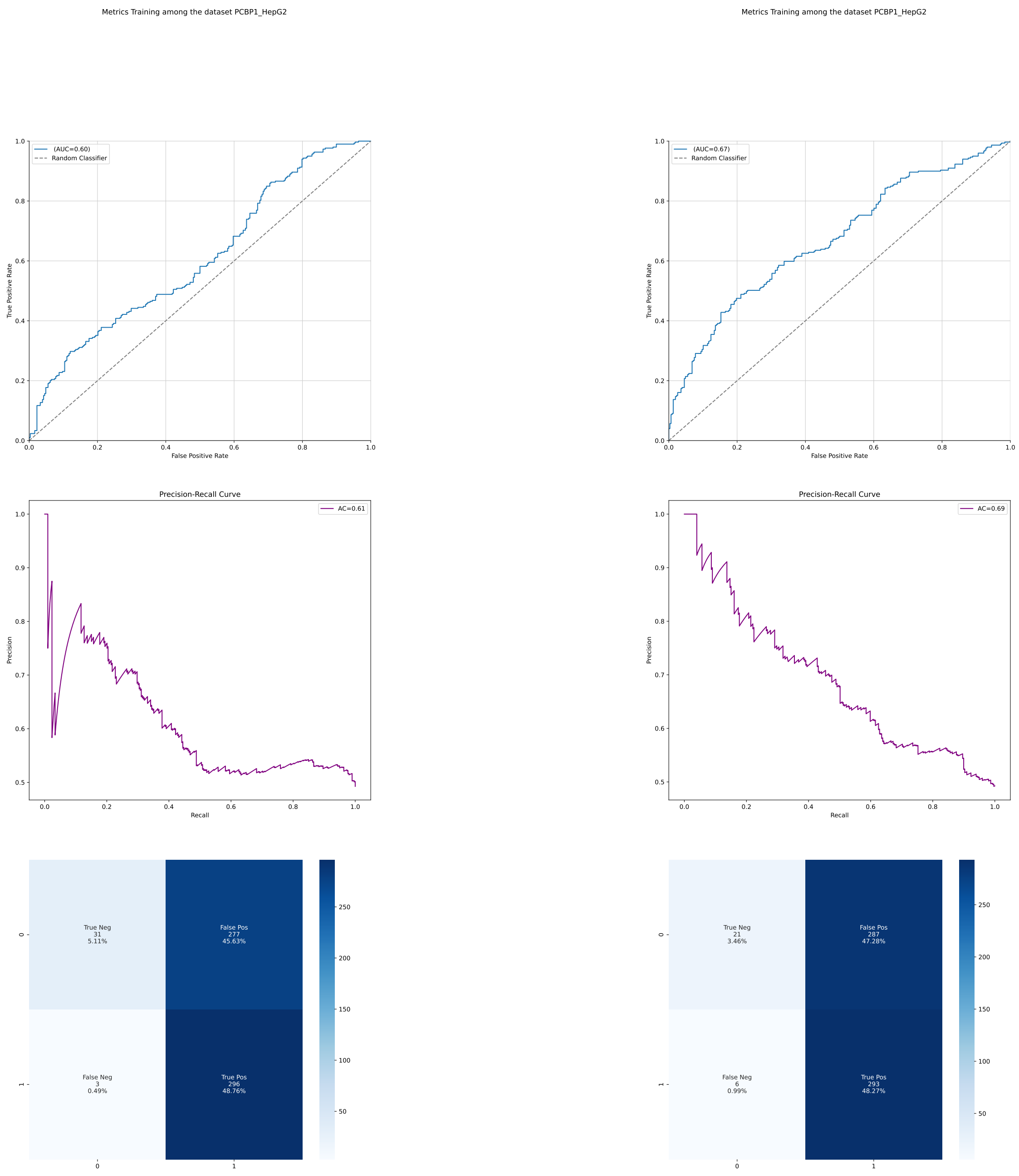
4.3.2. Negative 2

For the dataset containing only methylated sequences and utilizing the negative-2 set as the negative class, the obtained results are as follows.

The collected results encompass the accuracy, loss, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) of both the baseline and integrated models across training epochs. These results can be found in the Appendix A (Figure A.8). The observed curves strongly suggest the presence of overfitting, as evidenced by the loss values consistently below 0.1 and the accuracy metric approaching near-perfect performance (1.0). The label frequencies and size of the dataset used are to be found in Figure A.6.

Moreover, the analysis includes the evaluation of the area under the curve (AUC), precision-recall curve (PRC), and confusion matrix associated with the predictions, as depicted in Figure 4.13. Notably, the classification performance is comparable to the previous model that considered the negative-1 set as a negative class. The baseline model achieves an AUROC and AUPRC of 0.60 and 0.61, respectively, while the integrated model achieves an AUROC and AUPRC of 0.67 and 0.69, respectively. The baseline model achieves a correct classification rate of 53.87% for the sequences, whereas the integrated model achieves an accuracy of 51.73%. These results specifically pertain to the protein set PCBP1 from the HepG2 cell line.

However, the aggregated results presented in Table 4.7 provide a comprehensive discussion on the average AUROC and AUPRC across all investigated proteins. The findings shown in Figure 4.14 are consistent with the specific results, revealing an average receiver operating characteristic (ROC) score of 0.53 and an average precision-recall (PR) AUC of 0.54 for the baseline model. For the integrated model, the mean values for PR AUC and ROC AUC are 0.53 and 0.55, respectively.

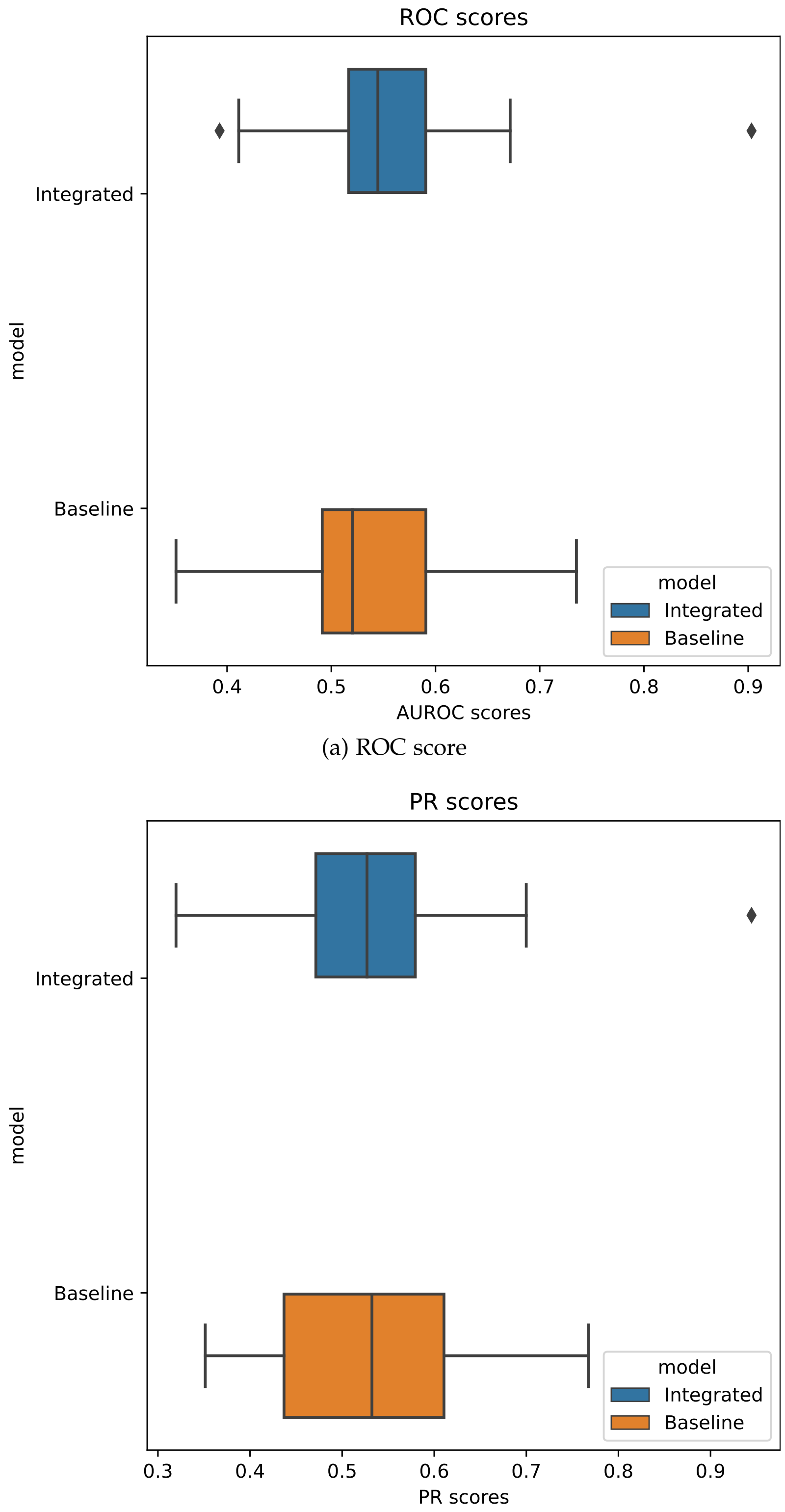


(a) Baseline Model

(b) Integrated Model

Figure 4.13.: Comparison of Baseline and Integrated Model: Prediction

Comparison of performance metrics for prediction between the baseline and integrated models: Area Under the Receiver Operating Characteristic Curve (AUROC) (top), Precision-Recall (PR) values (middle), and Confusion Matrix (bottom).



(a) ROC score

(b) PR score

Figure 4.14.: Comparison of Baseline and Integrated Model: Setting A negative-2

Table 4.7.: Aggregated Evaluation for Setting A - Negative2

Table 4.8.: Baseline Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.4864	0.5794
	DDX55	0.5488	0.5324
	DDX6	0.4542	0.3694
	IGF2BP3	0.5003	0.5559
	LARP4	0.5121	0.4858
	PCBP1	0.6039	0.6100
	RBM15	0.5203	0.5395
	SUB1	0.5000	0.4618
	UPF1	0.6829	0.7016
K562			
	AKAP1	0.7352	0.6845
	APOBEC3C	0.3510	0.4272
	CPEB4	0.4000	0.7675
	DDX55	0.5000	0.3768
	DDX6	0.4290	0.4427
	IGF2BP1	0.4415	0.6396
	IGF2BP2	0.4743	0.3828
	LARP4	0.5454	0.6112
	METAP2	0.5961	0.5973
	NOLC1	0.5275	0.5981
	NPM1	0.5443	0.4732
	PABPC4	0.4967	0.5173
	PCBP1	0.6513	0.6159
	PUM1	0.5855	0.4304
	PUM2	0.4962	0.4753
	RBM15	0.5982	0.5606
	RPS3	0.5032	0.3514
	SDAD1	0.6166	0.4311
	UPF1	0.5534	0.5240
	UTP3	0.4648	0.4072
	YBX3	0.5760	0.6819
	ZC3H11A	0.5965	0.6625

Table 4.9.: Integrated Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.5815	0.6346
	DDX55	0.5084	0.5270
	DDX6	0.5376	0.4834
	IGF2BP3	0.5117	0.5780
	LARP4	0.5157	0.5245
	PCBP1	0.6716	0.6878
	RBM15	0.5231	0.5209
	SUB1	0.4926	0.4788
	UPF1	0.5945	0.6999
K562			
	AKAP1	0.5177	0.4609
	APOBEC3C	0.5868	0.5967
	CPEB4	0.9032	0.9446
	DDX55	0.5471	0.4243
	DDX6	0.5619	0.5041
	IGF2BP1	0.4112	0.5704
	IGF2BP2	0.5451	0.4641
	LARP4	0.6083	0.6520
	METAP2	0.6441	0.6278
	NOLC1	0.5375	0.5760
	NPM1	0.3928	0.3196
	PABPC4	0.5064	0.4911
	PCBP1	0.6069	0.5364
	PUM1	0.6003	0.4399
	PUM2	0.5447	0.5326
	RBM15	0.6265	0.5809
	RPS3	0.5504	0.3832
	SDAD1	0.5327	0.3306
	UPF1	0.4499	0.4098
	UTP3	0.5244	0.5095
	YBX3	0.5219	0.5615
	ZC3H11A	0.5465	0.5770

4.4. Setting B

Considering the scarcity of methylated data in the protein sets, it was decided to only present the results for Setting B with the augmented dataset, not taking into consideration the dataset without augmented data. The results are divided as follows into *Negative 1* and *Negative 2* sections.

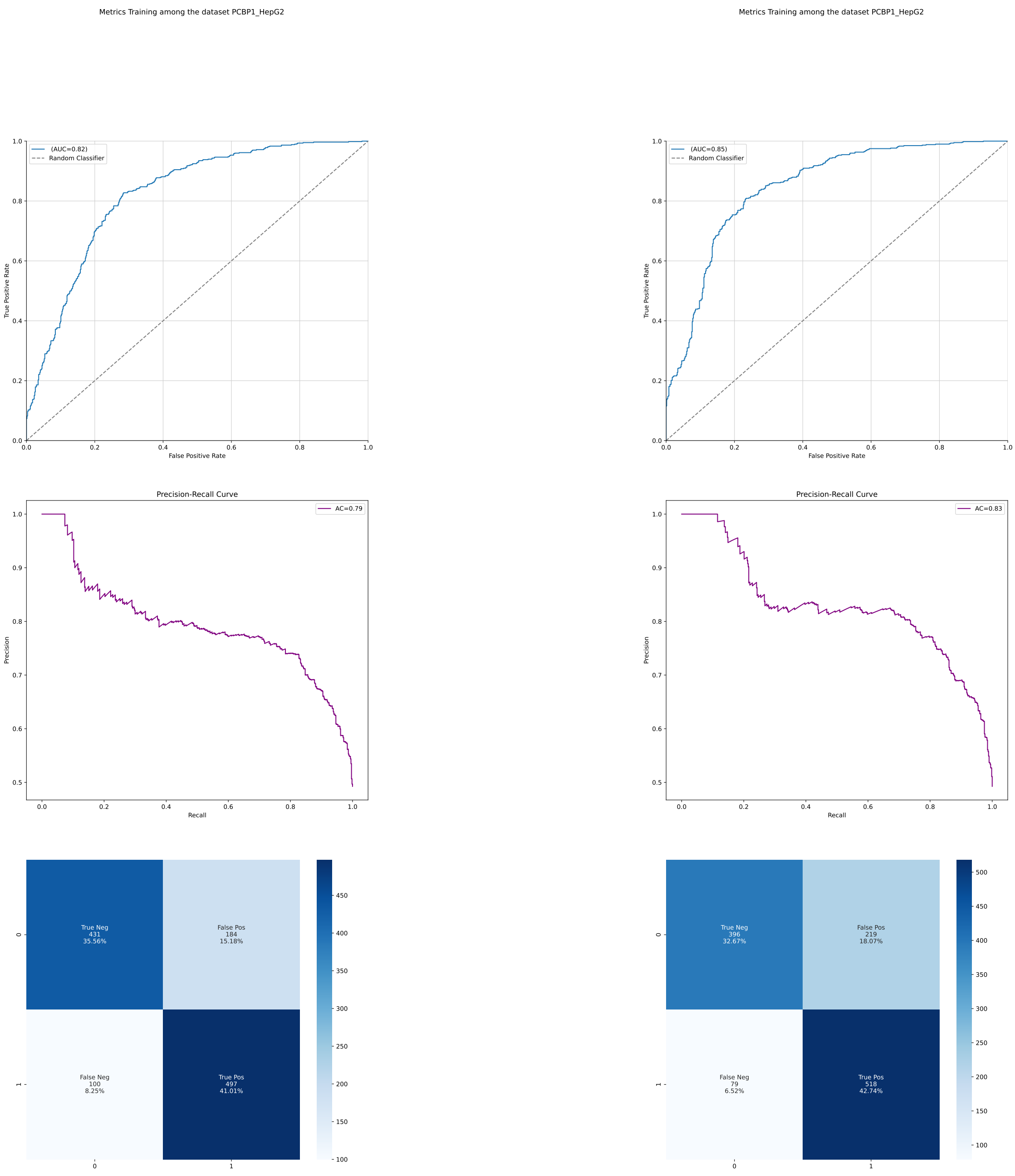
4.4.1. Negative 1

For the dataset containing a 1:1 ratio of non-methylated and methylated sequences, utilizing the negative-1 set as the negative class, whose label frequency is available in Figure A.9, the following results were obtained.

The collected results include the accuracy, loss, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) for both the baseline and integrated models across the training epochs. These results can be found in the Appendix A (Figure A.11). The observed curves indicate a loss value close to 0.1 and an accuracy of 0.95 for the baseline model, while the integrated model demonstrates a loss value of 0.25 and an accuracy of 0.9.

Furthermore, the analysis involves evaluating the area under the curve (AUC), precision-recall curve (PRC), and confusion matrix associated with the predictions, as shown in Figure 4.15. It is noteworthy that the classification performance improved compared to the previous model that considered only methylated sequences. The baseline model achieved an AUROC and AUPRC of 0.82 and 0.79, respectively, while the integrated model achieved an AUROC and AUPRC of 0.85 and 0.83, respectively. The baseline model achieved a correct classification rate of 76.57% for the sequences, while the integrated model achieved 75.41% accuracy. These results specifically pertain to the protein set PCBP1 from the HepG2 cell line.

However, the aggregated results, as depicted in their entirety in Table 4.10, provide a comprehensive discussion on the average AUROC and AUPRC across all investigated proteins. The findings presented in Figure 4.16 align with the specific results, displaying an average receiver operating characteristic (ROC) score of 0.73 and an average precision-recall (PR) AUC of 0.68 for the baseline model. As for the integrated model, the mean values for PR AUC and ROC AUC are 0.69 and 0.72, respectively.

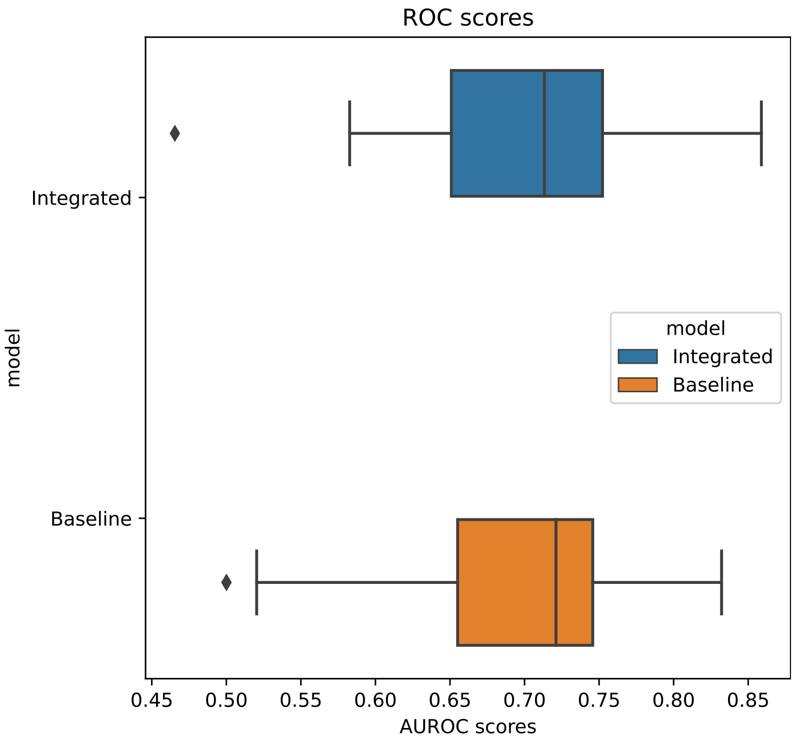


(a) Baseline Model

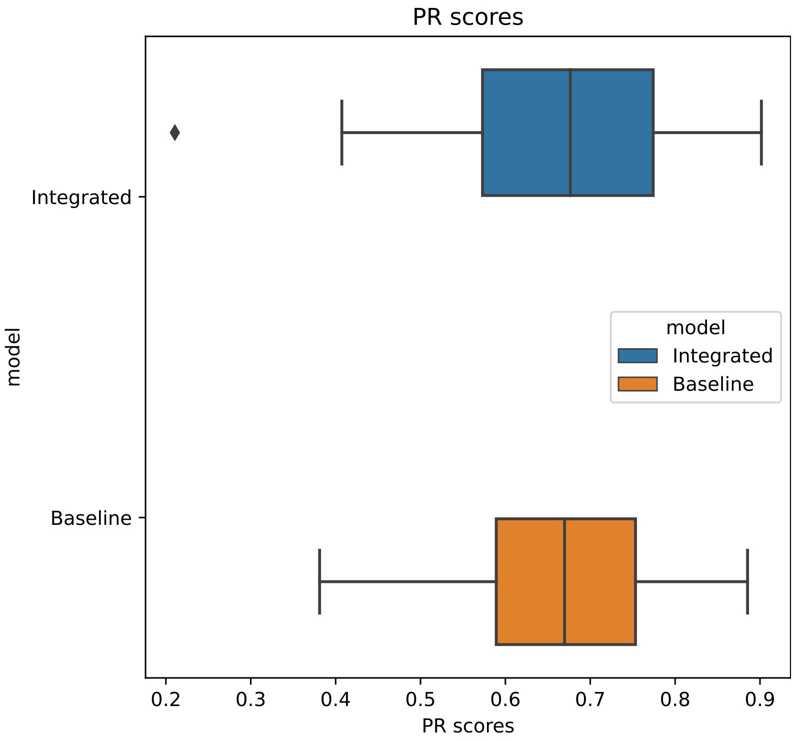
(b) Integrated Model

Figure 4.15.: Comparison of Baseline and Integrated Model: Prediction

Comparison of performance metrics for prediction between the baseline and integrated models: Area Under the Receiver Operating Characteristic Curve (AUROC) (top), Precision-Recall (PR) values (middle), and Confusion Matrix (bottom).



(a) ROC score



(b) PR score

Figure 4.16.: Comparison of Baseline and Integrated Model: Setting B negative-1

Table 4.10.: Aggregated Evaluation for Setting B - Negative1

Table 4.11.: Baseline Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.7333	0.7695
	DDX55	0.6442	0.5996
	DDX6	0.6966	0.5772
	IGF2BP3	0.7211	0.7436
	LARP4	0.6977	0.6681
	PCBP1	0.8216	0.7934
	RBM15	0.7306	0.7024
	SUB1	0.5000	0.4617
	UPF1	0.7515	0.7737
K562			
	AKAP1	0.8138	0.7769
	APOBEC3C	0.5203	0.5316
	CPEB4	0.7269	0.8853
	DDX55	0.6911	0.5711
	DDX6	0.7400	0.7231
	IGF2BP1	0.6512	0.7630
	IGF2BP2	0.7260	0.6273
	LARP4	0.6712	0.6697
	METAP2	0.8262	0.8198
	NOLC1	0.5001	0.5740
	NPM1	0.6742	0.3811
	PABPC4	0.6481	0.6137
	PCBP1	0.7998	0.7416
	PUM1	0.7706	0.6738
	PUM2	0.7323	0.7248
	RBM15	0.8322	0.8203
	RPS3	0.7285	0.5883
	SDAD1	0.8086	0.5691
	UPF1	0.6928	0.6385
	UTP3	0.6538	0.5902
	YBX3	0.6564	0.6738
	ZC3H11A	0.5705	0.6081

Table 4.12.: Integrated Model

Cell line	RBP	ROC AUC score	PR AUC score
HepG2			
	AKAP1	0.7315	0.7889
	DDX55	0.6364	0.6050
	DDX6	0.6624	0.5450
	IGF2BP3	0.6876	0.6955
	LARP4	0.6431	0.5836
	PCBP1	0.8469	0.8262
	RBM15	0.7221	0.6582
	SUB1	0.5961	0.5063
	UPF1	0.7635	0.7721
K562			
	AKAP1	0.7792	0.7465
	APOBEC3C	0.6514	0.6807
	CPEB4	0.7822	0.9017
	DDX55	0.6598	0.5309
	DDX6	0.7180	0.6766
	IGF2BP1	0.5827	0.6916
	IGF2BP2	0.7128	0.5464
	LARP4	0.7411	0.7762
	METAP2	0.8149	0.7962
	NOLC1	0.6441	0.7162
	NPM1	0.4655	0.2106
	PABPC4	0.6566	0.5625
	PCBP1	0.8589	0.7905
	PUM1	0.7665	0.6531
	PUM2	0.7145	0.6037
	RBM15	0.7872	0.7762
	RPS3	0.7133	0.5470
	SDAD1	0.7194	0.4073
	UPF1	0.7069	0.5925
	UTP3	0.6448	0.6380
	YBX3	0.6504	0.6783
	ZC3H11A	0.7231	0.7799

4.4.2. Negative 2

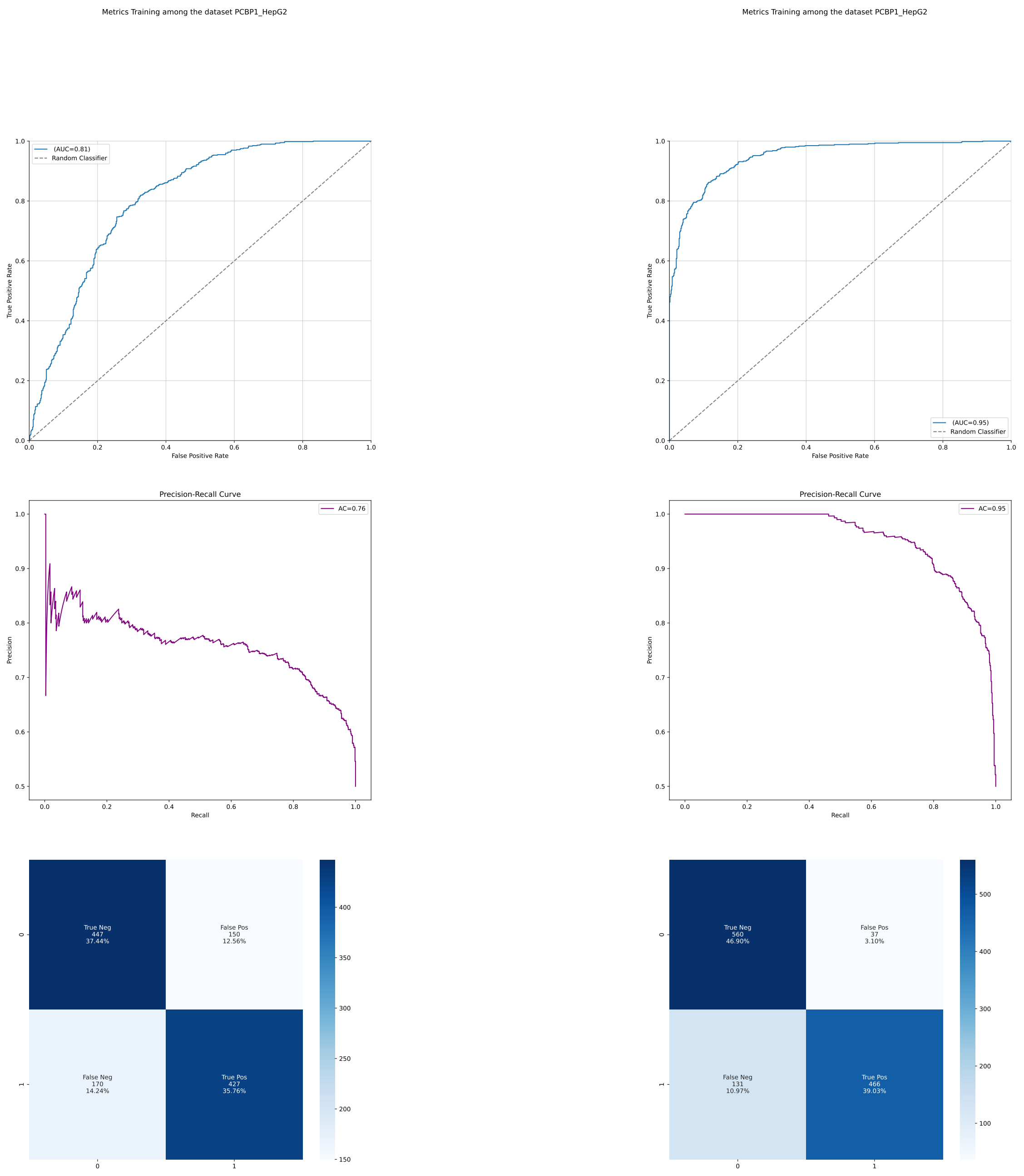
The obtained results are for the dataset containing an equal ratio of non-methylated and methylated sequences, with the negative-2 set serving as the negative class.

As a reference, the label frequencies and dataset size can be found in Figure A.10; in this case the negatives, particularly numerous, have been downsampled to match the number of positives and enable a balanced dataset. The collected results include various metrics including accuracy, loss, area under the receiver operating characteristic curve (AUROC), and area under the precision-recall curve (AUPRC) for both the baseline and integrated models across the training epochs. Detailed results can be found in Appendix A (Figure A.12). Analysis of the observed curves indicates a loss value of 0.3 and an accuracy of 0.8 for the baseline model, while the integrated model exhibits a loss value of 0.4 and an accuracy of

0.85.

Furthermore, the evaluation of the predictions involves examining the area under the curve (AUC), precision-recall curve (PRC), and confusion matrix. These aspects are depicted in Figure 4.17. Notably, the classification performance demonstrates improvement compared to the previous model that considered only methylated sequences. The baseline model achieved an AUROC and AUPRC of 0.81 and 0.76, respectively, whereas the integrated model achieved an AUROC and AUPRC of 0.95 and 0.95, respectively. The baseline model achieved a correct classification rate of 73.2% for the sequences, while the integrated model achieved an accuracy of 85.93%. These results pertain specifically to the protein set PCBP1 from the HepG2 cell line.

However, when examining the aggregated results presented in their entirety in Table 4.13, a comprehensive discussion on the average AUROC and AUPRC across all investigated proteins can be derived. The findings depicted in Figure 4.18 align with the specific results, indicating an average receiver operating characteristic (ROC) score of 0.75 and an average precision-recall (PR) AUC of 0.73 for the baseline model. In the case of the integrated model, the mean values for PR AUC and ROC AUC are 0.92 and 0.90, respectively.

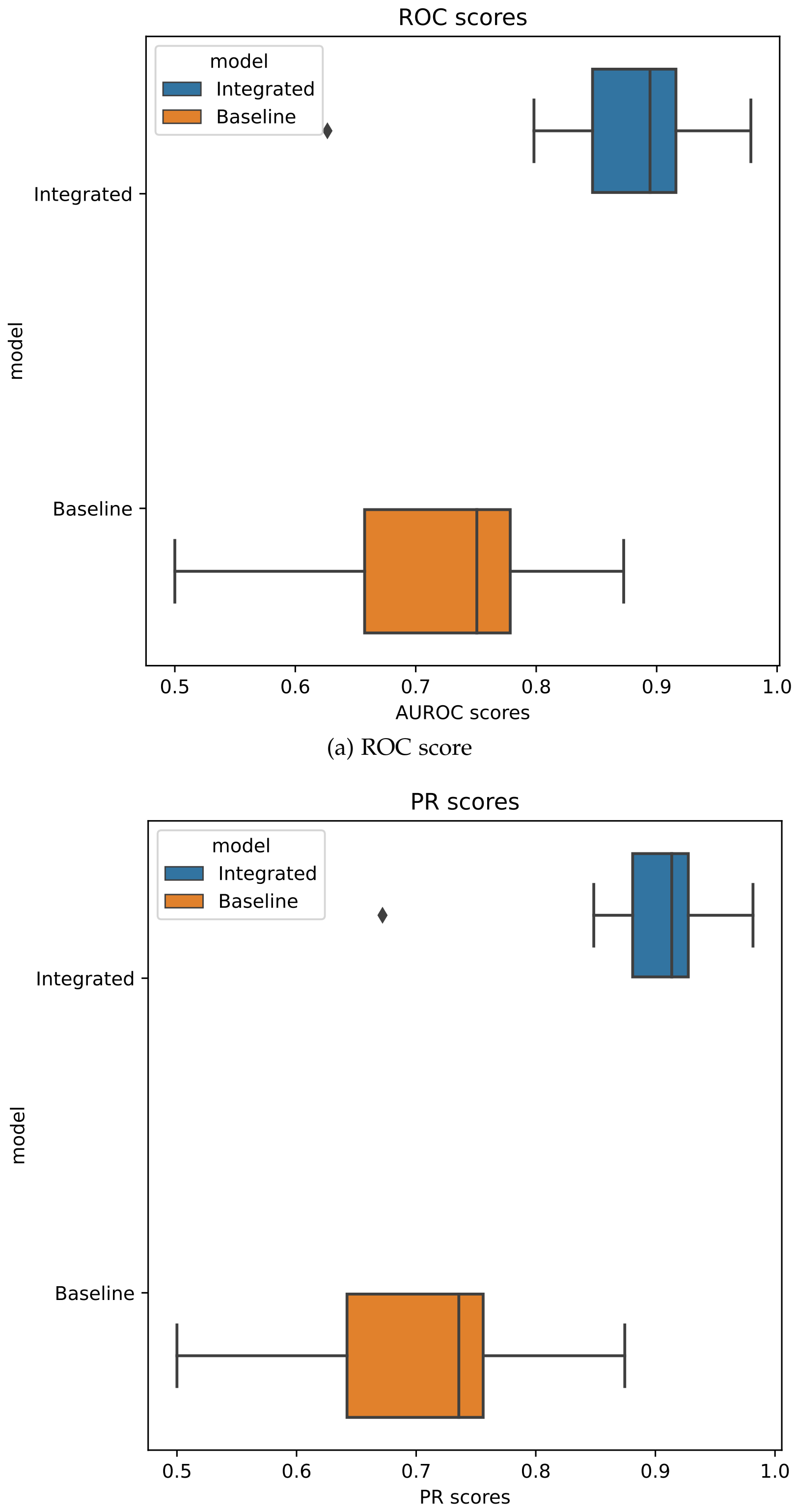


(a) Baseline Model

(b) Integrated Model

Figure 4.17.: Comparison of Baseline and Integrated Model: Prediction

Comparison of performance metrics for prediction between the baseline and integrated models: Area Under the Receiver Operating Characteristic Curve (AUROC) (top), Precision-Recall (PR) values (middle), and Confusion Matrix (bottom).



(a) ROC score

(b) PR score

Figure 4.18.: Comparison of Baseline and Integrated Model: Setting B negative-2

Table 4.13.: Aggregated Evaluation for Setting B - Negative2

Table 4.14.: Baseline Model

Cell line	RBP	roc$_{a}uc_{s}core$	pr$_{a}uc_{s}core$
HepG2	AKAP1	0.8001	0.7801
HepG2	DDX55	0.6973	0.6804
HepG2	DDX6	0.5000	0.5000
HepG2	IGF2BP3	0.7619	0.7412
HepG2	LARP4	0.5002	0.5002
HepG2	PCBP1	0.8066	0.7672
HepG2	RBM15	0.7507	0.7316
HepG2	SUB1	0.7395	0.7289
HepG2	UPF1	0.7909	0.7574
K562	AKAP1	0.7639	0.7428
K562	APOBEC3C	0.6371	0.6296
K562	CPEB4	0.7517	0.7718
K562	DDX55	0.6741	0.6426
K562	DDX6	0.5000	0.5000
K562	IGF2BP1	0.8169	0.7914
K562	IGF2BP2	0.7697	0.7542
K562	LARP4	0.6363	0.6417
K562	METAP2	0.7719	0.7546
K562	NOLC1	0.7019	0.6821
K562	NPM1	0.5607	0.5675
K562	PABPC4	0.7613	0.7525
K562	PCBP1	0.8295	0.8033
K562	PUM1	0.7692	0.7422
K562	PUM2	0.8726	0.8744
K562	RBM15	0.8335	0.8171
K562	RPS3	0.7504	0.7120
K562	SDAD1	0.7336	0.7356
K562	UPF1	0.5000	0.5000
K562	UTP3	0.6411	0.6195
K562	YBX3	0.7850	0.7534
K562	ZC3H11A	0.6890	0.6731

Table 4.15.: Integrated Model

Cell line	RBP	roc$_{a}uc_{s}core$	pr$_{a}uc_{s}core$
HepG2	AKAP1	0.8966	0.9137
HepG2	DDX55	0.8385	0.8752
HepG2	DDX6	0.7981	0.8485
HepG2	IGF2BP3	0.9245	0.9338
HepG2	LARP4	0.8372	0.8780
HepG2	PCBP1	0.9504	0.9540
HepG2	RBM15	0.8933	0.9137
HepG2	SUB1	0.8957	0.9139
HepG2	UPF1	0.9155	0.9273
K562	AKAP1	0.9021	0.9182
K562	APOBEC3C	0.8348	0.8718
K562	CPEB4	0.8963	0.9176
K562	DDX55	0.8492	0.8840
K562	DDX6	0.8243	0.8694
K562	IGF2BP1	0.9165	0.9277
K562	IGF2BP2	0.9182	0.9305
K562	LARP4	0.8377	0.8769
K562	METAP2	0.9020	0.9206
K562	NOLC1	0.8665	0.8954
K562	NPM1	0.6267	0.6719
K562	PABPC4	0.8945	0.9166
K562	PCBP1	0.9646	0.9671
K562	PUM1	0.8913	0.9122
K562	PUM2	0.9782	0.9816
K562	RBM15	0.9462	0.9529
K562	RPS3	0.8801	0.9001
K562	SDAD1	0.9320	0.9360
K562	UPF1	0.8812	0.9049
K562	UTP3	0.8443	0.8722
K562	YBX3	0.8985	0.9150
K562	ZC3H11A	0.8622	0.8945

5. Discussion

The aim of this research is to investigate the impact of N(6)-methyladenosine (m6A) modifications on RNA-protein interactions and develop integrated models that combine RNA sequence and m6A data to predict the binding preferences of RNA-binding proteins (RBPs). By doing so, the study aims to gain insights into the regulatory code governing RNA-protein interactions and understand the interplay between m6A modifications and RBPs in the context of RNA processing and gene expression.

The integration of m6A data into the models could allow for a comprehensive understanding of the cooperation between m6A modifications and RBPs in RNA-protein interactions. By exploring whether there is a correlation between the binding affinity of RBPs and the presence of m6A modifications in the vicinity of the binding sites, the research contributes to our understanding of the intricate mechanisms that govern gene expression and post-transcriptional regulation. These findings have significant implications for our comprehension of the functional consequences of m6A methylations and their role in shaping RNA-protein interactions. By investigating the effects of m6A modifications on RNA-protein interactions, the study attempts to shed light on the regulatory mechanisms underlying gene expression and could possibly provide insights into how dysregulation of m6A modifications or RBPs can lead to pathological conditions and diseases [5, 6, 7, 8]. Moreover, the development of integrated models combining RNA sequence and m6A data represents a notable advancement in the field. These models have the potential to accurately predict RBP binding preferences, serving as valuable tools for studying RNA-protein interactions and facilitating future investigations. The insights gained from this research could have repercussions for the development of therapeutic strategies targeting RBPs and RNA modifications, offering the potential for novel treatments aimed at diseases associated with dysregulated gene expression. In summary, this research not only aims to identify the impact of m6A modifications on RNA-protein interactions but also contributes to our broader understanding of the regulatory mechanisms governing gene expression. The findings have implications for the functional consequences of m6A methylations, shed light on the interplay between m6A modifications and RBPs, and provide a foundation for the development of predictive models and therapeutic strategies. By elucidating the complex regulatory networks involved in RNA biology, this research opens new avenues for advancing our understanding of cellular processes and disease mechanisms.

The study obtained several important findings regarding the prediction of RBP binding in the presence of methylated sequences. It was observed that relying solely on a dataset consisting of methylated sequences for predicting RBP binding did not yield satisfactory results. However, when both methylated and non-methylated sequences were introduced in equal proportions, the predictive models showed satisfactory performance by successfully distinguishing between the positive class (bound sequences) and the negative class (unbound sequences or sequences bound to other RBPs). Additionally, when utilizing the negative-2 as negative set and accounting for the impact of methylation data, in Setting B, the classification performance of the models improved. This finding aligns with the hypothesis that m6A sites have an influence on the binding preferences of RBPs. It has been demonstrated that m6A modifications control the accessibility of RNA binding motifs (RBMs) by influencing the RNA structure, thereby impacting RNA-protein interactions and biological regulation. This regulatory mechanism is referred to as the "m6A-switch." The study revealed that m6A modifications induce local structural changes in both mRNA and long non-coding RNA (lncRNA), facilitating the binding of an abundant nuclear RNA-binding protein involved in pre-mRNA processing. The concept of the "m6A-switch" lines up with previous research, suggesting that m6A modifications act as a molecular switch, altering the three-dimensional structure of RNA and enhancing the affinity for RBP binding [75]. These findings provide a mechanistic understanding of how m6A methylations influence RNA-protein interactions and support the hypothesis that m6A modifications play a crucial role in determining RBP binding preferences. This observation reinforces the notion that considering the impact of m6A modifications is essential for accurate prediction of RBP binding. The enhanced predictive power achieved by incorporating methylation data further underscores the significance of m6A modifications in shaping RBP binding preferences and their potential as regulatory elements in post-transcriptional gene regulation.

The preliminary analysis identified several RBPs which are known to be associated with m6A sites. Specifically, the family IGF2BP (IGF2BP1, IGF2BP2, IGF2BP3) exhibited high methylation rates, consistent with their known association with m6A sites [12]. Other RBP families implicated in m6A binding, such as the YTH family and the HNRNP family, did not show high methylation rates or significant m6A site counts (**??**). Moreover, the maximum percentage of m6A occurrence was observed in PABPC4 (K562 cell line) at 20% Figure 4.5. Notably, PABPC4 has been linked to m6A mRNA expression and is considered required for the promotion of the expression of at least two genes [76], supporting the results of the preliminary analysis. In 4.1a, EXOSC5 (K562 cell line) is depicted as an outlier positioned in the top left corner of the graph. Although the percentage appears to be zero, it is actually 0.13, indicating the presence of a considerable average number of m6A sites per methylated sequence, despite the low methylation rate. During the preliminary analysis, it was observed that the number of methylated sequences present in each protein set was notably low (Figure 4.5). This scarcity of m6A site data poses a challenge for training an effective binary classifier, as a sufficient amount of data is necessary to achieve better generalization, reduce bias, and optimize performance. It is expected that m6A sites occur at a frequency of only 0.15-0.6% of all adenosines [77] and to address this data scarcity issue, protein sets with relatively higher methylation rates were selected for this study. This selection strategy is particularly relevant for the Setting A dataset, which exclusively contains methylated sequences, aiming to mitigate the impact of data scarcity.

Upon evaluating the results of Setting A for the augmented data (see Figure 4.12 and Figure 4.11), it became evident that the data augmentation process did not yield satisfactory outcomes. This finding was unexpected and contrary to the hypothesis that the integration of m6A sites would improve the classification performance of protein binding. The observed opposite results were not in line with the anticipated hypothesis. Consequently, further analysis was conducted to investigate the reliability of these unforeseen findings and identify potential technical issues that may have influenced the results.

An important observation was that not only did the performance of the integrated model drop, but the performance of the baseline models also exhibited a decline. This raised the critical question of whether this decline was attributed to the size of the training set or, possibly, the unsuccessful nature of the data augmentation process. To address this question, a comparison was made between the performances of the baseline models trained on two datasets with equivalent sizes, one containing the augmented dataset and the other containing sequences produced from the original binding sites coordinates [69]. The results, as depicted in Figure 4.10, revealed that the downsized model's performance for the examined protein PCBP1 (HepG2 cell line) was unexpectedly high, considering the reduced number of data points inputted, and aligned with the results of the model trained with all the expressed genes. This finding indicates that the drop in performance observed in Setting A was not solely due to the dataset size, but rather due to the limitations of the employed data augmentation method.

The hypothesis is that the augmentation method is not suitable for this particular model, possibly due to the shifting of the sequences; this alone might not have effectively increased the variance within the dataset. To accurately reflect real-world conditions, the dataset should encompass a wide range of variations and scenarios. However, the employed augmentation approach in Setting A merely augmented the number of data points without substantially improving the model's training. In the case of a binary classifier like the one used in this study, a well-sampled dataset containing representative data is essential. The presence of irrelevant or noisy data can hinder the model's performance and introduce unnecessary complexity.

Moreover, upon careful examination of the performance and metrics outlined in chapter 4 (Figure 4.5), it was observed that there is no apparent correlation between the protein sets that exhibited enhanced performance following the integration of methylated data, as depicted in Table A.4, and the number of m6A sites present in each respective protein set. This suggests that the influence of the methylated data on model performance is not solely determined by the abundance of m6A sites, indicating the presence of additional factors influencing the model's behaviour and highlighting the complexity of the relationship between m6A sites and protein binding. Therefore, it is recommended to revisit Setting A and explore alternative data augmentation methods that can effectively increase the variance within the dataset. This may involve considering other types of sequence modifications or incorporating additional features to capture a wider range of variations. By doing so, it is expected that the model's performance can be enhanced, leading to more accurate predictions.

The results of Setting B were also unexpected. The relevance of the failed augmentation for Setting B, which did not appear to have a significant impact, can be attributed to the composition of the dataset. In this setting, half of the dataset consists of augmented data, while the other half comprises non-augmented data lacking m6A sites. Surprisingly, this dataset composition seems to provide sufficient variance and relevance for the model to effectively classify sequences during the prediction phase. The initial hypothesis was that the model's performance would be adversely affected by the 1:1 ratio of methylated and non-methylated sequences, resulting in lower performance compared to Setting A. However, this was not the case, suggesting the presence of another bias in the model's behaviour. It is possible that the model does not specifically recognize the m6A sites but rather the binding motif for the examined protein. However, this hypothesis is contradicted by the results. If the model solely recognized the binding motif, there would be no improvement in the integrated model compared to the baseline model, which lacks the additional information encoded in the 5th channel representing the m6A sites. As shown in Figure Figure 4.18, this is not the case for the negative-2 set; instead, the integration of the m6A information enhances the model's performance.

However, the same improvement is not observed for the set containing negative-1 as the negative class (Figure Figure 4.16). The discrepancy in performance between negative-1 and negative-2 sets may be attributed to the variance within the datasets. Both negative-1 and negative-2 sets were augmented at a ratio of 1:30, but the negative-2 set had a higher occurrence of m6A sites in its sequences. Additionally, negative-2, although classified as a negative class in this study, consists of sequences bound to proteins. Previous studies[78, 72], have reported a higher occurrence of m6A sites in sequences that bind to RNA-binding proteins (RBPs) compared to those lacking binding motifs for RBPs. Interestingly, the findings align with this study, which demonstrated a strong positive correlation between the number of m6A sites, miRNAs, and RBPs binding to mRNAs. This suggests that m6A-modified mRNAs are more likely to be targeted by RBPs. Furthermore, the study revealed that m6A sites are located proximally to binding sites of multiple RBPs.

To address this issue, a downsampling technique was applied to the negative2 set in Setting B, as mentioned in subsection 4.4.2. By downsampling, the original negative-2 dataset containing m6A sites was reduced to achieve a balanced dataset. Consequently, the higher variance present in the original negative-2 dataset, as well as the reduction of repeated subsequences due to downsampling, likely contributed to the improved classification performance of the model. Furthermore, concerns were raised regarding the possibility of the model learning specific subsequences from the augmented data and relying solely on them for predictions. However, the results indicate that the model is capable of recognizing not only the augmented data but also the non-augmented data, which may contain smaller amounts or lack the repeated subsequences present in the augmented data.

In conclusion, the preliminary analysis conducted in this study provides valuable insights into the relationship between m6A sites and RBPs. However, it is important to acknowledge the limitations and potential impact of the study design and procedure on the obtained results. Due to time restraints, cross-validation was not implemented, and the preparation of the m6A-filtered dataset was particularly time-consuming. These factors limit the robustness of the results and suggest the need for further validation through cross-validation in future investigations. Additionally, individual tuning of the model for each RBP should be explored to assess the potential improvement in performance.

Furthermore, the scarcity of m6A site data poses a challenge for training an effective binary classifier. The frequency of m6A sites is relatively low, ranging from 0.15% to 0.6% of all adenosines. To mitigate the impact of data scarcity, protein sets with relatively higher methylation rates were selected in this study. However, it is worth noting that the selection of proteins based on methylation rates depends on the size of the initial dataset and may not always reflect the actual number of m6A sites, as seen in 4.1b. As demonstrated in Figure 4.2, not all datasets contain the same number of data points, making the reliance on methylation rate less relevant. Therefore, future investigations could consider exploring alternative selection criteria, such as focusing on proteins with the highest absolute number of m6A sites.

In addition to addressing the limitations of the current study, future investigations should also consider the use of alternative m6A site datasets. Although miCLIP is the state-of-the-art method for in-vivo transcriptome-wide identification of m6A sites, it is not the only dataset available and may not offer the highest level of accuracy. For example, the MAZTER-seq [79] dataset provides an alternative approach to m6A site detection, albeit with limited coverage. This dataset specifically identifies m6A sites associated with a specific nucleotide sequence ('ACA'), which may result in an incomplete representation of all m6A sites present in the transcriptome. Exploring the MAZTER-seq dataset in future investigations could provide complementary insights and help validate the findings obtained using miCLIP. Another alternative dataset worth considering is GLORI[43], which offers an antibody-free method for m6A site detection with improved specificity and resolution. Compared to miCLIP, GLORI demonstrates higher specificity and resolution in identifying m6A sites. Although miCLIP was initially selected in this study due to its single nucleotide resolution and widespread adoption within the research community, exploring the GLORI dataset could provide valuable corroborating evidence and enhance our understanding of m6A site and RBP interactions.

Additionally, the choice of encoding method is an important aspect to consider in future investigations. While one-hot encoding was employed for both the sequences and methylation data in this study, alternative encoding methods should be explored, especially for the methylation data. One potential approach is to encode the overall level of methylation within a sequence and introduce it as separate information alongside the one-hot encoded sequence. This can be achieved by incorporating additional neurons in the model dedicated to capturing this supplementary information [14]. Such an approach could offer additional insights into the impact of methylation levels on post-transcriptional regulation and further enhance the model's capacity to learn from the methylation data.

In summary, this study serves as a stepping stone towards understanding the intricate relationship between m6A sites and RBPs. The findings underscore the importance of implementing cross-validation, individual tuning for RBPs, developing improved data augmentation methods, exploring alternative m6A site datasets, and considering alternative encoding methods. By addressing these aspects in future research, we can enhance the predictive models, deepen our understanding of post-transcriptional regulation, and contribute to advancements in the field of RNA biology.

6. Conclusion

In conclusion, this thesis has explored and investigated the intricate relationship between N(6)-methyladenosine modifications and RNA-protein interactions, aiming to unravel the underlying mechanisms and shed light on the regulatory role of m6A in gene expression. The findings of this research provide valuable insights into the impact of m6A modifications on RNA-protein interactions and contribute to our broader understanding of the regulatory code governing post-transcriptional gene regulation. Through the integration of m6A data into deep learning models, this thesis has exemplified the significance of considering m6A methylations in predicting RNA-binding protein binding preferences. The observed correlation between m6A modifications and the binding affinity of RBPs supports the notion that m6A plays a crucial role in shaping RNA-protein interactions. The concept of the "m6A-switch" mechanism, where m6A alters the RNA three-dimensional structure and enhances the affinity for RBP binding, has been supported by the results of this research. These findings provide a mechanistic perception of how m6A methylations influence RNA-protein interactions and contribute to the regulation of gene expression. However, this thesis has also highlighted certain limitations and challenges associated with the integration of m6A data into predictive models. The scarcity of m6A site data and the need for well-sampled datasets containing representative data pose significant obstacles in training effective binary classifiers. The findings underline the importance of exploring alternative data augmentation methods to increase the variance within the dataset and improve model performance. Additionally, the selection of examined proteins based on methylation rates should be reconsidered, and alternative criteria, such as the absolute number of m6A sites, should be inspected. Future research should focus on addressing the constraints identified in this thesis and further validating the uncoverings. Cross-validation techniques should be implemented to reinforce the robustness of the results, and the individual tuning of the model for each RBP could potentially improve the performance and provide more accurate predictions. Moreover, the utilization of alternative m6A site datasets, such as MAZTER-seq and GLORI, could be investigated to validate and complement the findings obtained using miCLIP. These additional datasets may offer improved accuracy, specificity, and resolution in identifying m6A sites. Furthermore, future investigations could consider alternative encoding methods, particularly for the methylation data. Exploring different approaches, such as incorporating the overall methylation level as separate information from the sequence, could provide additional insights into the impact of methylation levels on post-transcriptional regulation. Overall, this thesis has advanced our understanding of the interplay between m6A modifications and RBPs in RNA-protein interactions. The findings have implications for deciphering the regulatory mechanisms governing gene expression and could potentially contribute to the development of therapeutic strategies targeting RBPs and RNA modifications. By unravelling the intricate regulatory networks involved in RNA biology, this thesis opens new avenues for advancing our understanding of cellular processes and disease mechanisms.

Bibliography

[1] A. Q. Gomes, S. Nolasco, and H. Soares. “Non-coding RNAs: multi-tasking molecules in the cell”. In: *International journal of molecular sciences* 14.8 (2013), pp. 16010–16039.

[2] J. E. Wilusz and P. A. Sharp. “A circuitous route to noncoding RNA”. In: *Science* 340.6131 (2013), pp. 440–441.

[3] S. H. Boo and Y. K. Kim. “The emerging role of RNA modifications in the regulation of mRNA stability”. In: *Experimental & molecular medicine* 52.3 (2020), pp. 400–408.

[4] T. Glisovic, J. L. Bachorik, J. Yong, and G. Dreyfuss. “RNA-binding proteins and post-transcriptional gene regulation”. In: *FEBS letters* 582.14 (2008), pp. 1977–1986.

[5] J. D. Keene. “RNA regulons: coordination of post-transcriptional events”. In: *Nature Reviews Genetics* 8.7 (2007), pp. 533–543.

[6] E. Dassi. “Handshakes and fights: the regulatory interplay of RNA-binding proteins”. In: *Frontiers in molecular biosciences* 4 (2017), p. 67.

[7] L. Wurth and F. Gebauer. “RNA-binding proteins, multifaceted translational regulators in cancer”. In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1849.7 (2015), pp. 881–886.

[8] L. De Conti, M. Baralle, and E. Buratti. “Neurodegeneration and RNA-binding proteins”. In: *Wiley Interdisciplinary Reviews: RNA* 8.2 (2017), e1394.

[9] K. Kapeli, F. J. Martinez, and G. W. Yeo. “Genetic mutations in RNA-binding proteins and their roles in ALS”. In: *Human genetics* 136 (2017), pp. 1193–1214.

[10] A. Ramakrishnan and S. C. Janga. “Human protein-RNA interaction network is highly stable across mammals”. In: *BMC genomics* 20 (2019), pp. 1–14.

[11] N. Liu, K. I. Zhou, M. Parisien, Q. Dai, L. Diatchenko, and T. Pan. “N6-methyladenosine alters RNA structure to regulate binding of a low-complexity protein”. In: *Nucleic acids research* 45.10 (2017), pp. 6051–6063.

[12] Y. Zhao, Y. Shi, H. Shen, and W. Xie. “m6A-binding proteins: the emerging crucial performers in epigenetics”. In: *Journal of hematology & oncology* 13.1 (2020), pp. 1–14.

[13] M. Horlacher, G. Cantini, J. Hesse, P. Schinke, N. Goedert, S. Londhe, L. Moyon, and A. Marsico. “A Systematic Benchmark of Machine Learning Methods for Protein-RNA Interaction Prediction”. In: *bioRxiv* (2023), pp. 2023–02.

[14] S. Budach. “Dissertation: Explainable Deep Learning Models For Biological Sequence Classification”. In: (2019).

[15] L. R. Ganser, M. L. Kelly, D. Herschlag, and H. M. Al-Hashimi. "The roles of structural dynamics in the cellular functions of RNAs". In: *Nature reviews Molecular cell biology* 20.8 (2019), pp. 474–489.

[16] H. Chen and B. F. Pugh. "What do transcription factors interact with?" In: *Journal of molecular biology* 433.14 (2021), p. 166883.

[17] S. A. Lambert, A. Jolma, L. F. Campitelli, P. K. Das, Y. Yin, M. Albu, X. Chen, J. Taipale, T. R. Hughes, and M. T. Weirauch. "The human transcription factors". In: *Cell* 172.4 (2018), pp. 650–665.

[18] A. Fradera-Sola, E. Nischwitz, M. E. Bayer, K. Luck, and F. Butter. "RNA-dependent interactome allows network-based assignment of RNA-binding protein function". In: *Nucleic Acids Research* (2023), gkad245.

[19] M. Corley, M. C. Burns, and G. W. Yeo. "How RNA-binding proteins interact with RNA: molecules and mechanisms". In: *Molecular cell* 78.1 (2020), pp. 9–29.

[20] S. Gerstberger, M. Hafner, and T. Tuschl. "A census of human RNA-binding proteins". In: *Nature Reviews Genetics* 15.12 (2014), pp. 829–845.

[21] A. Jolma, J. Zhang, E. Mondragón, E. Morgunova, T. Kivioja, K. U. Laverty, Y. Yin, F. Zhu, G. Bourenkov, Q. Morris, et al. "Binding specificities of human RNA-binding proteins toward structured and linear RNA sequences". In: *Genome research* 30.7 (2020), pp. 962–973.

[22] F. E. Baralle and J. Giudice. "Alternative splicing as a regulator of development and tissue identity". In: *Nature reviews Molecular cell biology* 18.7 (2017), pp. 437–451.

[23] A. Hasan, C. Cotobal, C. D. Duncan, and J. Mata. "Systematic analysis of the role of RNA-binding proteins in the regulation of RNA stability". In: *PLoS genetics* 10.11 (2014), e1004684.

[24] S. M. Garcıa-Mauriño, F. Rivero-Rodrıguez, A. Velázquez-Cruz, M. Hernández-Vellisca, A. Dıaz-Quintana, M. A. De la Rosa, and I. Dıaz-Moreno. "RNA binding protein regulation and cross-talk in the control of AU-rich mRNA fate". In: *Frontiers in molecular biosciences* 4 (2017), p. 71.

[25] R. F. Harvey, T. S. Smith, T. Mulroney, R. M. Queiroz, M. Pizzinga, V. Dezi, E. Villenueva, M. Ramakrishna, K. S. Lilley, and A. E. Willis. "Trans-acting translational regulatory RNA binding proteins". In: *Wiley Interdisciplinary Reviews: RNA* 9.3 (2018), e1465.

[26] W. Arif, G. Datar, and A. Kalsotra. "Intersections of post-transcriptional gene regulatory mechanisms with intermediary metabolism". In: *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1860.3 (2017), pp. 349–362.

[27] N. Jonkhout, J. Tran, M. A. Smith, N. Schonrock, J. S. Mattick, and E. M. Novoa. "The RNA modification landscape in human disease". In: *Rna* 23.12 (2017), pp. 1754–1769.

[28] R. V. Kadumuri and S. C. Janga. "Epitranscriptomic code and its alterations in human disease". In: *Trends in Molecular Medicine* 24.10 (2018), pp. 886–903.

[29] I. A. Roundtree, M. E. Evans, T. Pan, and C. He. "Dynamic RNA modifications in gene expression regulation". In: *Cell* 169.7 (2017), pp. 1187–1200.

[30] S. Oerum, V. Meynier, M. Catala, and C. Tisné. "A comprehensive review of m6A/m6Am RNA methyltransferase structures". In: *Nucleic acids research* 49.13 (2021), pp. 7239–7255.

[31] B. Linder, A. V. Grozhik, A. O. Olarerin-George, C. Meydan, C. E. Mason, and S. R. Jaffrey. "Single-nucleotide-resolution mapping of m6A and m6Am throughout the transcriptome". In: *Nature methods* 12.8 (2015), pp. 767–772.

[32] S. Wang, W. Lv, T. Li, S. Zhang, H. Wang, X. Li, L. Wang, D. Ma, Y. Zang, J. Shen, et al. "Dynamic regulation and functions of mRNA m6A modification". In: *Cancer cell international* 22.1 (2022), p. 48.

[33] Z. Wang, M. Gerstein, and M. Snyder. "RNA-Seq: a revolutionary tool for transcriptomics". In: *Nature reviews genetics* 10.1 (2009), pp. 57–63.

[34] M. F. Rai, E. D. Tycksen, L. J. Sandell, and R. H. Brophy. "Advantages of RNA-seq compared to RNA microarrays for transcriptome profiling of anterior cruciate ligament tears". In: *Journal of Orthopaedic Research®* 36.1 (2018), pp. 484–497.

[35] E. L. Van Nostrand, G. A. Pratt, B. A. Yee, E. C. Wheeler, S. M. Blue, J. Mueller, S. S. Park, K. E. Garcia, C. Gelboin-Burkhart, T. B. Nguyen, et al. "Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins". In: *Genome biology* 21.1 (2020), pp. 1–26.

[36] M. Hafner, M. Katsantoni, T. Köster, J. Marks, J. Mukherjee, D. Staiger, J. Ule, and M. Zavolan. "CLIP and complementary methods". In: *Nature Reviews Methods Primers* 1.1 (2021), p. 20.

[37] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule. "iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution". In: *Nature structural & molecular biology* 17.7 (2010), pp. 909–915.

[38] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, et al. "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)". In: *Nature methods* 13.6 (2016), pp. 508–514.

[39] B. J. Zarnegar, R. A. Flynn, Y. Shen, B. T. Do, H. Y. Chang, and P. A. Khavari. "irCLIP platform for efficient characterization of protein–RNA interactions". In: *Nature methods* 13.6 (2016), pp. 489–492.

[40] T. Ni, V. Majerciak, Z.-M. Zheng, and J. Zhu. "PA-seq for Global Identification of RNA Polyadenylation Sites of Kaposi's Sarcoma–Associated Herpesvirus Transcripts". In: *Current protocols in microbiology* 41.1 (2016), 14E–7.

[41] D. Dierks, M. A. Garcia-Campos, A. Uzonyi, M. Safra, S. Edelheit, A. Rossi, T. Sideri, R. A. Varier, A. Brandis, Y. Stelzer, et al. "Multiplexed profiling facilitates robust m6A quantification at site, gene and sample resolution". In: *Nature methods* 18.9 (2021), pp. 1060–1067.

[42] B. Molinie, J. Wang, K. S. Lim, R. Hillebrand, Z.-x. Lu, N. Van Wittenberghe, B. D. Howard, K. Daneshvar, A. C. Mullen, P. Dedon, et al. “m6A-LAIC-seq reveals the census and complexity of the m6A epitranscriptome”. In: *Nature methods* 13.8 (2016), pp. 692–698.

[43] C. Liu, H. Sun, Y. Yi, W. Shen, K. Li, Y. Xiao, F. Li, Y. Li, Y. Hou, B. Lu, et al. “Absolute quantification of single-base m6A methylation in the mammalian transcriptome using GLORI”. In: *Nature Biotechnology* 41.3 (2023), pp. 355–366.

[44] H. Shin, T. Liu, X. Duan, Y. Zhang, and X. S. Liu. “Computational methodology for ChIP-seq analysis”. In: *Quantitative biology* 1 (2013), pp. 54–70.

[45] T. Bailey, P. Krajewski, I. Ladunga, C. Lefebvre, Q. Li, T. Liu, P. Madrigal, C. Taslim, and J. Zhang. “Practical guidelines for the comprehensive analysis of ChIP-seq data”. In: *PLoS computational biology* 9.11 (2013), e1003326.

[46] G. Yu, L.-G. Wang, and Q.-Y. He. “ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization”. In: *Bioinformatics* 31.14 (2015), pp. 2382–2383.

[47] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, et al. “Model-based analysis of ChIP-Seq (MACS)”. In: *Genome biology* 9.9 (2008), pp. 1–9.

[48] S. Krakau, H. Richard, and A. Marsico. “PureCLIP: capturing target-specific protein–RNA interaction footprints from single-nucleotide CLIP-seq data”. In: *Genome biology* 18 (2017), pp. 1–17.

[49] P. J. Uren, E. Bahrami-Samani, S. C. Burns, M. Qiao, F. V. Karginov, E. Hodges, G. J. Hannon, J. R. Sanford, L. O. Penalva, and A. D. Smith. “Site identification in high-throughput RNA–protein interaction data”. In: *Bioinformatics* 28.23 (2012), pp. 3013–3020.

[50] X. Li, H. Kazan, H. D. Lipshitz, and Q. D. Morris. “Finding the target sites of RNA-binding proteins”. In: *Wiley Interdisciplinary Reviews: RNA* 5.1 (2014), pp. 111–130.

[51] J. Fostier. “BLAMM: BLAS-based algorithm for finding position weight matrix occurrences in DNA sequences on CPUs and GPUs”. In: *BMC bioinformatics* 21 (2020), pp. 1–13.

[52] M. K. Das and H.-K. Dai. “A survey of DNA motif finding algorithms”. In: *BMC bioinformatics* 8.7 (2007), pp. 1–13.

[53] T. L. Bailey, N. Williams, C. Misleh, and W. W. Li. “MEME: discovering and analyzing DNA and protein sequence motifs”. In: *Nucleic acids research* 34.suppl_2 (2006), W369–W373.

[54] N. T. T. Nguyen, B. Contreras-Moreira, J. A. Castro-Mondragon, W. Santana-Garcia, R. Ossio, C. D. Robles-Espinoza, M. Bahin, S. Collombet, P. Vincens, D. Thieffry, et al. “RSAT 2018: regulatory sequence analysis tools 20th anniversary”. In: *Nucleic acids research* 46.W1 (2018), W209–W214.

[55] Y. Fan, M. Kon, and C. DeLisi. “Transcription factor-DNA binding via machine learning ensembles”. In: *arXiv preprint arXiv:1805.03771* (2018).

[56] O. A. Montesinos López, A. Montesinos López, and J. Crossa. “Fundamentals of Artificial Neural Networks and Deep Learning”. In: *Multivariate Statistical Machine Learning Methods for Genomic Prediction*. Springer, 2022, pp. 379–425.

[57] S. Walczak and N. Cerpa. “Heuristic principles for the design of artificial neural networks”. In: *Information and software technology* 41.2 (1999), pp. 107–117.

[58] O. I. Abiodun, A. Jantan, A. E. Omolara, K. V. Dada, N. A. Mohamed, and H. Arshad. “State-of-the-art in artificial neural network applications: A survey”. In: *Heliyon* 4.11 (2018), e00938.

[59] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey. “Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning”. In: *Nature biotechnology* 33.8 (2015), pp. 831–838.

[60] J. Zhou and O. G. Troyanskaya. “Predicting effects of noncoding variants with deep learning–based sequence model”. In: *Nature methods* 12.10 (2015), pp. 931–934.

[61] D. R. Kelley, J. Snoek, and J. L. Rinn. “Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks”. In: *Genome research* 26.7 (2016), pp. 990–999.

[62] B. Lee, J. Baek, S. Park, and S. Yoon. “deepTarget: end-to-end learning framework for microRNA target prediction using deep recurrent neural networks”. In: *Proceedings of the 7th ACM international conference on bioinformatics, computational biology, and health informatics*. 2016, pp. 434–442.

[63] S. Park, S. Min, H. Choi, and S. Yoon. “deepMiRGene: Deep neural network based precursor microrna prediction”. In: *arXiv preprint arXiv:1605.00017* (2016).

[64] V. Boža, B. Brejová, and T. Vinař. “DeepNano: deep recurrent neural networks for base calling in MinION nanopore reads”. In: *PloS one* 12.6 (2017), e0178751.

[65] S. Singh, Y. Yang, B. Póczos, and J. Ma. “Predicting enhancer-promoter interaction from genomic sequence with deep neural networks”. In: *Quantitative Biology* 7 (2019), pp. 122–137.

[66] P. Mineault and K. Nozawa. *patrickmineault/codebook: 1.0.0*. Version releases. Dec. 2021. DOI: `10.5281/zenodo.5796873`. URL: `https://doi.org/10.5281/zenodo.5796873`.

[67] Z. Sun, S. Xue, H. Xu, X. Hu, S. Chen, Z. Yang, Y. Yang, J. Ouyang, and H. Cui. “Effects of NSUN2 deficiency on the mRNA 5-methylcytosine modification and gene expression profile in HEK293 cells”. In: *Epigenomics* 11.4 (2019), pp. 439–453.

[68] E. P. Consortium et al. “An integrated encyclopedia of DNA elements in the human genome”. In: *Nature* 489.7414 (2012), p. 57.

[69] E. L. Van Nostrand, P. Freese, G. A. Pratt, X. Wang, X. Wei, R. Xiao, S. M. Blue, J.-Y. Chen, N. A. Cody, D. Dominguez, et al. “A large-scale binding and functional map of human RNA-binding proteins”. In: *Nature* 583.7818 (2020), pp. 711–719.

[70] B. T. Lee, G. P. Barber, A. Benet-Pagès, J. Casper, H. Clawson, M. Diekhans, C. Fischer, J. N. Gonzalez, A. S. Hinrichs, C. M. Lee, et al. "The UCSC genome browser database: 2022 update". In: *Nucleic Acids Research* 50.D1 (2022), pp. D1115–D1122.

[71] S. Ke, E. A. Alemu, C. Mertens, E. C. Gantman, J. J. Fak, A. Mele, B. Haripal, I. Zucker-Scharff, M. J. Moore, C. Y. Park, et al. "A majority of m6A residues are in the last exons, allowing the potential for 3 UTR regulation". In: *Genes & development* 29.19 (2015), pp. 2037–2053.

[72] W. Hong, Y. Zhao, Y.-L. Weng, and C. Cheng. "Random Forest model reveals the interaction between N6-methyladenosine modifications and RNA-binding proteins". In: *Iscience* 26.3 (2023).

[73] F. Chollet et al. *Keras*. 2015. URL: `https://github.com/fchollet/keras`.

[74] S. Budach and A. Marsico. "Pysster: classification of biological sequences by learning sequence and structure motifs with convolutional neural networks". In: *Bioinformatics* 34.17 (2018), pp. 3035–3037.

[75] N. Liu, Q. Dai, G. Zheng, C. He, M. Parisien, and T. Pan. "N 6-methyladenosine-dependent RNA structural switches regulate RNA–protein interactions". In: *Nature* 518.7540 (2015), pp. 560–564.

[76] D. A. Kuppers, S. Arora, Y. Lim, A. R. Lim, L. M. Carter, P. D. Corrin, C. L. Plaisier, R. Basom, J. J. Delrow, S. Wang, et al. "N6-methyladenosine mRNA marking promotes selective translation of regulons required for human erythropoiesis". In: *Nature communications* 10.1 (2019), p. 4596.

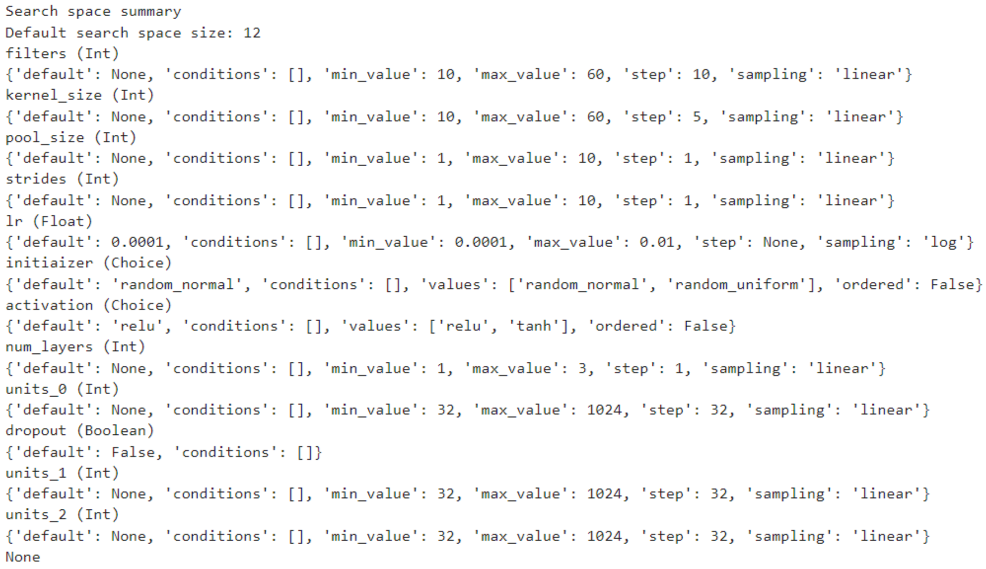
[77] P. C. He and C. He. "m6A RNA methylation: from mechanisms to therapeutic potential". In: *The EMBO journal* 40.3 (2021), e105977.

[78] S. D. Mandal and P. S. Ray. "Transcriptome-wide analysis reveals spatial correlation between N6-methyladenosine and binding sites of microRNAs and RNA-binding proteins". In: *Genomics* 113.1 (2021), pp. 205–216.

[79] M. A. Garcia-Campos, S. Edelheit, U. Toth, M. Safra, R. Shachar, S. Viukov, R. Winkler, R. Nir, L. Lasman, A. Brandis, et al. "Deciphering the "m6A code" via antibody-independent quantitative profiling". In: *Cell* 178.3 (2019), pp. 731–747.

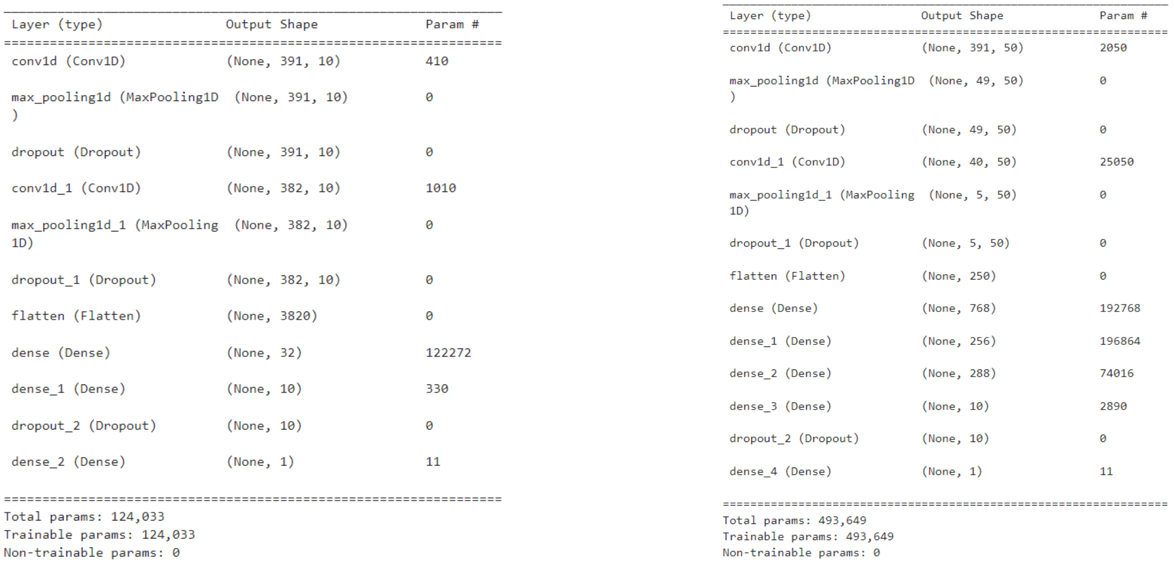
A. Appendix

A.1. Materials and Methods



```
Search space summary
Default search space size: 12
filters (Int)
{'default': None, 'conditions': [], 'min_value': 10, 'max_value': 60, 'step': 10, 'sampling': 'linear'}
kernel_size (Int)
{'default': None, 'conditions': [], 'min_value': 10, 'max_value': 60, 'step': 5, 'sampling': 'linear'}
pool_size (Int)
{'default': None, 'conditions': [], 'min_value': 1, 'max_value': 10, 'step': 1, 'sampling': 'linear'}
strides (Int)
{'default': None, 'conditions': [], 'min_value': 1, 'max_value': 10, 'step': 1, 'sampling': 'linear'}
lr (Float)
{'default': 0.0001, 'conditions': [], 'min_value': 0.0001, 'max_value': 0.01, 'step': None, 'sampling': 'log'}
initiaizer (Choice)
{'default': 'random_normal', 'conditions': [], 'values': ['random_normal', 'random_uniform'], 'ordered': False}
activation (Choice)
{'default': 'relu', 'conditions': [], 'values': ['relu', 'tanh'], 'ordered': False}
num_layers (Int)
{'default': None, 'conditions': [], 'min_value': 1, 'max_value': 3, 'step': 1, 'sampling': 'linear'}
units_0 (Int)
{'default': None, 'conditions': [], 'min_value': 32, 'max_value': 1024, 'step': 32, 'sampling': 'linear'}
dropout (Boolean)
{'default': False, 'conditions': []}
units_1 (Int)
{'default': None, 'conditions': [], 'min_value': 32, 'max_value': 1024, 'step': 32, 'sampling': 'linear'}
units_2 (Int)
{'default': None, 'conditions': [], 'min_value': 32, 'max_value': 1024, 'step': 32, 'sampling': 'linear'}
None
```

Figure A.1.: Hyperparameter tuning search space



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 391, 10)           410

 max_pooling1d (MaxPooling1D  (None, 391, 10)          0
 )

 dropout (Dropout)           (None, 391, 10)           0

 conv1d_1 (Conv1D)           (None, 382, 10)           1010

 max_pooling1d_1 (MaxPooling  (None, 382, 10)          0
 1D)

 dropout_1 (Dropout)         (None, 382, 10)           0

 flatten (Flatten)           (None, 3820)              0

 dense (Dense)               (None, 32)                122272

 dense_1 (Dense)             (None, 10)                330

 dropout_2 (Dropout)         (None, 10)                0

 dense_2 (Dense)             (None, 1)                 11

=================================================================
Total params: 124,033
Trainable params: 124,033
Non-trainable params: 0
```

(a) Untuned Model Architecture

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv1d (Conv1D)             (None, 391, 50)           2050

 max_pooling1d (MaxPooling1D  (None, 49, 50)           0
 )

 dropout (Dropout)           (None, 49, 50)            0

 conv1d_1 (Conv1D)           (None, 40, 50)            25050

 max_pooling1d_1 (MaxPooling  (None, 5, 50)            0
 1D)

 dropout_1 (Dropout)         (None, 5, 50)             0

 flatten (Flatten)           (None, 250)               0

 dense (Dense)               (None, 768)               192768

 dense_1 (Dense)             (None, 256)               196864

 dense_2 (Dense)             (None, 288)               74016

 dense_3 (Dense)             (None, 10)                2890

 dropout_2 (Dropout)         (None, 10)                0

 dense_4 (Dense)             (None, 1)                 11

=================================================================
Total params: 493,649
Trainable params: 493,649
Non-trainable params: 0
```

(b) Suggested Tuned Model Architecture

Figure A.2.: Architecture of the Untuned and Tuned Models

Parameter	Value
name	Adam
weight decay	None
clipnorm	None
global clipnorm	None
clipvalue	None
use ema	False
ema momentum	0.99
ema overwrite frequency	None
jit compile	False
is legacy optimizer	False
learning rate	0.0007744959
beta 1	0.9
beta 2	0.999
epsilon	1e07
amsgrad	False

Table A.1.: Optimizer configuration for the suggested Tuned Model

Hyperparameter	Value
filters	50
kernel_size	10
pool_size	10
strides	8
learning_rate	0.0007744959
initializer	′random_uniform′
activation	′tanh′
num_layers	3
units_0	768
dropout	False
units_1	256
units_2	288

Table A.2.: Hyperparameters for the suggested Tuned Model

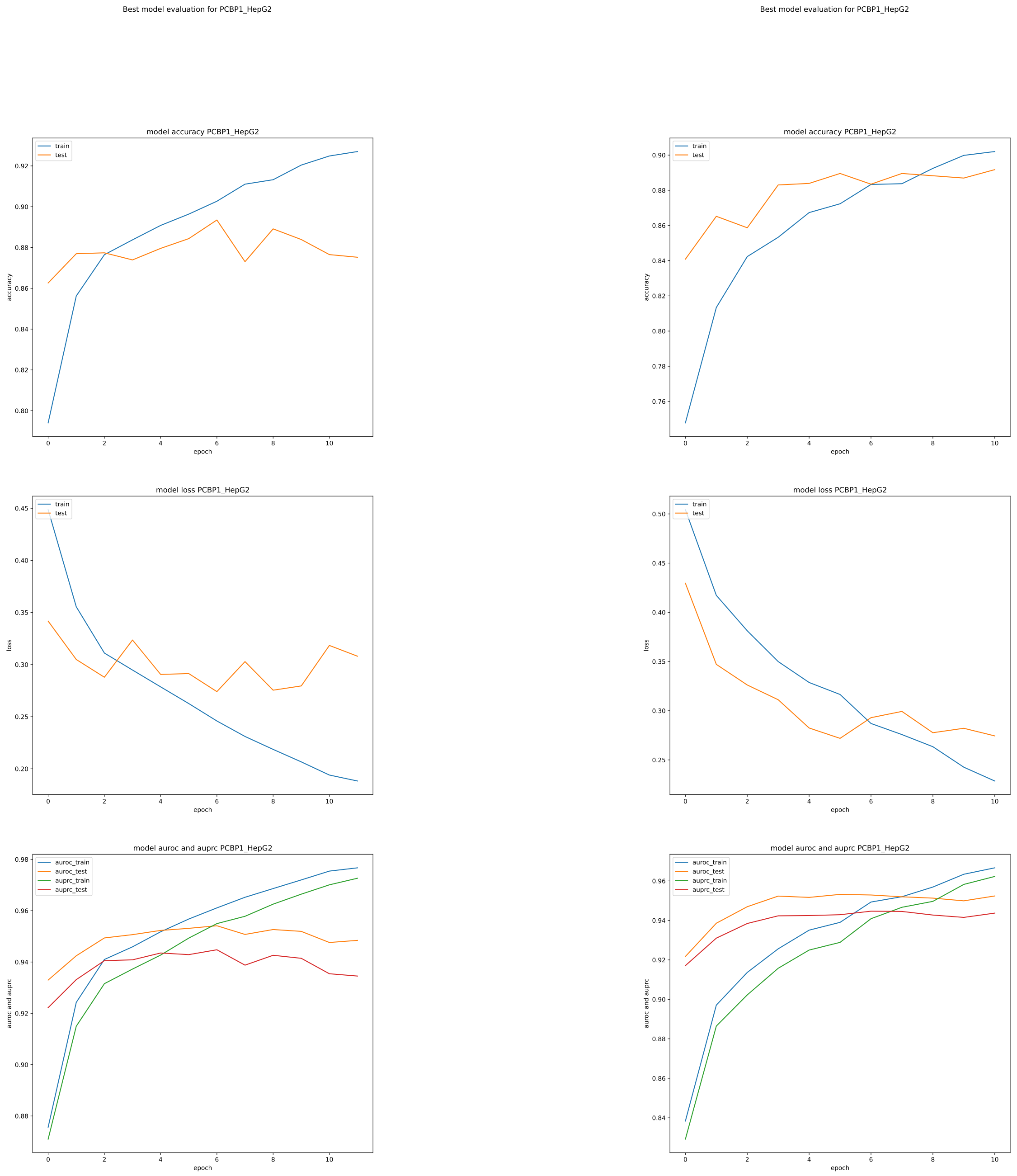
A.2. Results

A.2.1. Preprocessing

RBP	percentage
AKAP1_HepG2	14.2734
AKAP1_K562	11.2134
APOBEC3C_K562	18.0462
CPEB4_K562	12.5310
DDX55_HepG2	14.8182
DDX55_K562	15.5983
DDX6_HepG2	10.5101
DDX6_K562	12.2585
IGF2BP1_K562	12.3952
IGF2BP2_K562	12.8726
IGF2BP3_HepG2	13.4018
LARP4_HepG2	18.8550
LARP4_K562	12.6007
METAP2_K562	12.8981
NOLC1_K562	13.7495
NPM1_K562	10.3529
PABPC4_K562	20.0167
PCBP1_HepG2	11.2069
PCBP1_K562	10.8336
PUM1_K562	10.7035
PUM2_K562	11.2132
RBM15_HepG2	12.0292
RBM15_K562	15.1247
RPS3_K562	12.4834
SDAD1_K562	10.8808
SUB1_HepG2	16.9396
UPF1_HepG2	14.3785
UPF1_K562	10.4727
UTP3_K562	10.3789
YBX3_K562	14.1573
ZC3H11A_K562	12.5471

Table A.3.: Caption

A.2.2. Tuning of the Model



(a) Untuned Model

(b) Tuned Model

Figure A.3.: Comparison of Standard and Tuned Model: Training

Comparison of performance metrics during training for the untuned and tuned models: Accuracy (top), Loss (middle), and Area Under the Receiver Operating Characteristic Curve (AUROC) and Precision-Recall (PR) values (bottom) plotted over the training epochs.

A.2.3. Setting A

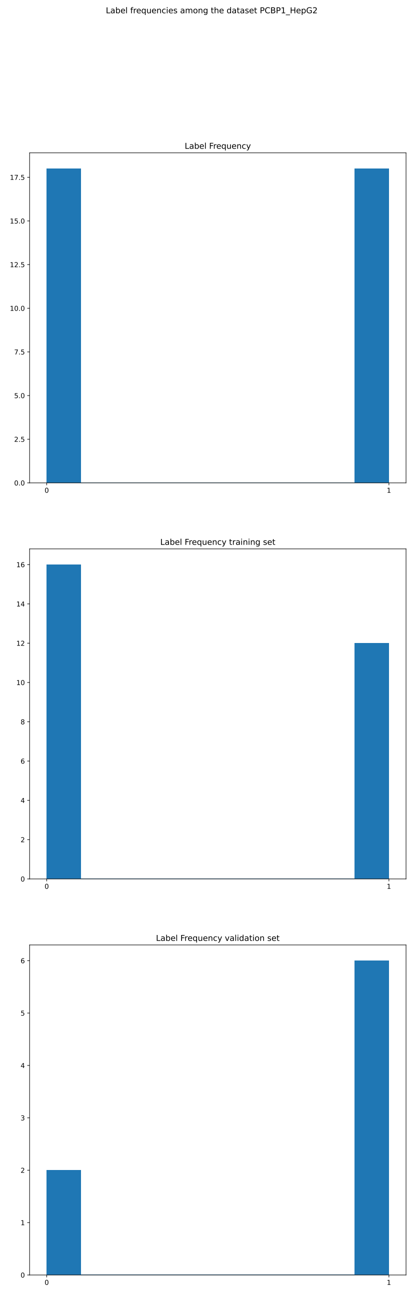


Figure A.4.: Label frequency for Setting A dataset Negative 1

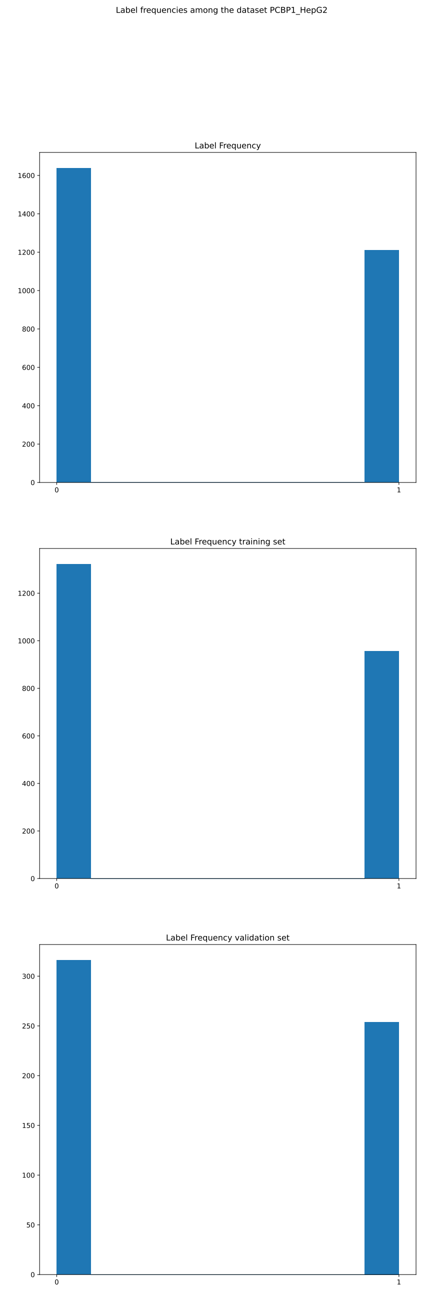
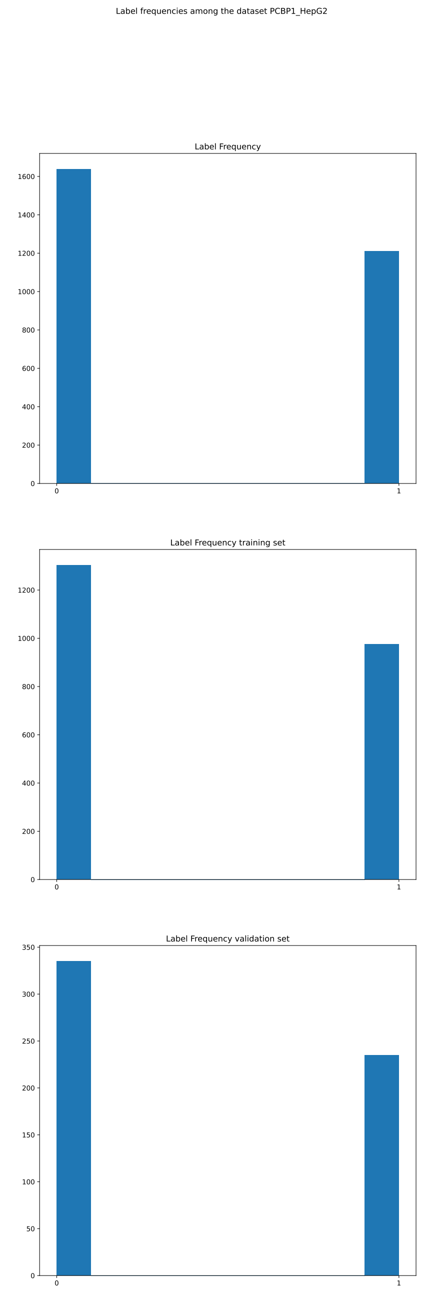


Figure A.5.: Label frequency for Setting A Augmented dataset Negative 1



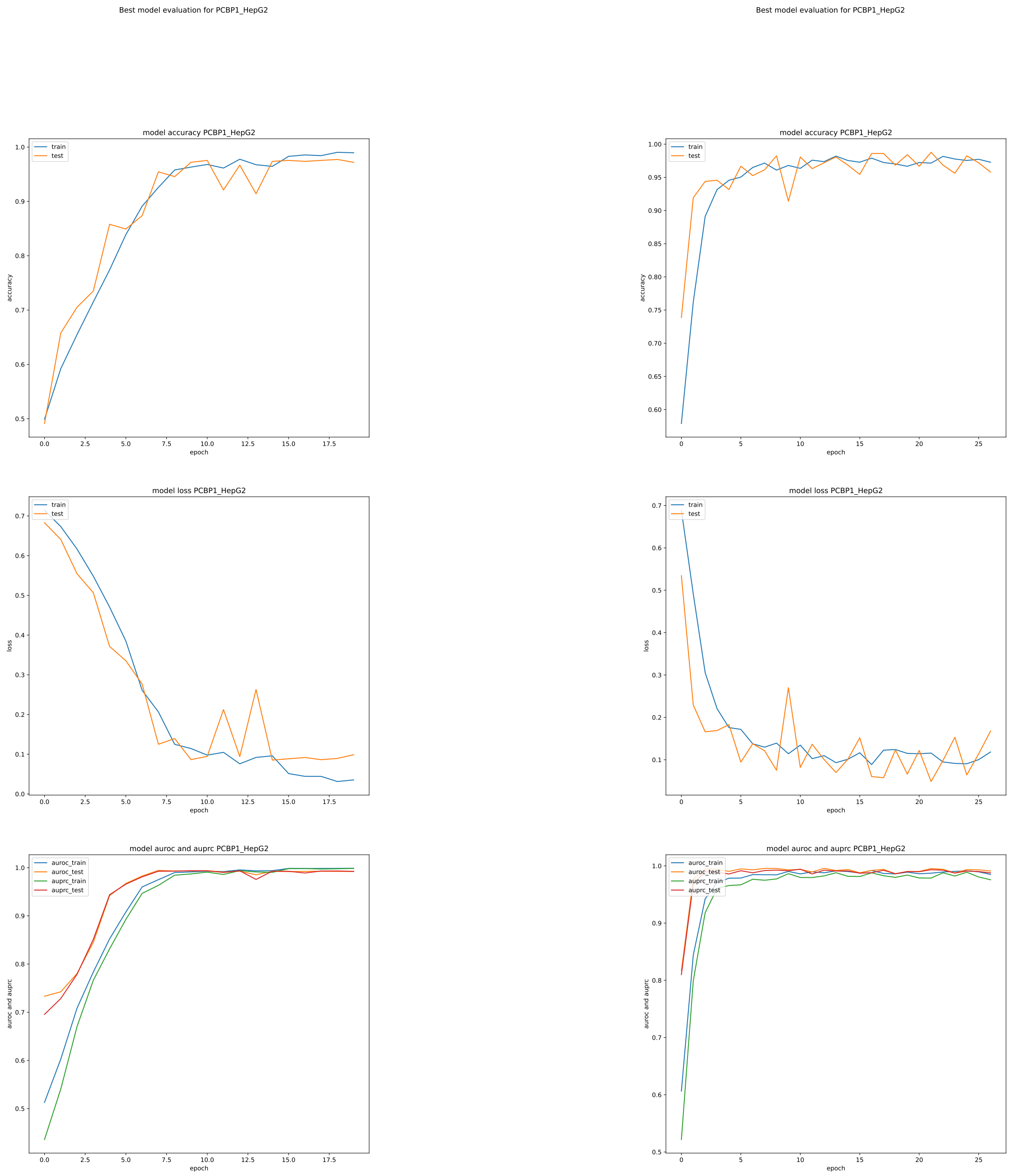
Label frequencies among the dataset PCBP1_HepG2

Label Frequency

Label Frequency training set

Label Frequency validation set

Figure A.6.: Label frequency for Setting A Augmented dataset Negative 2

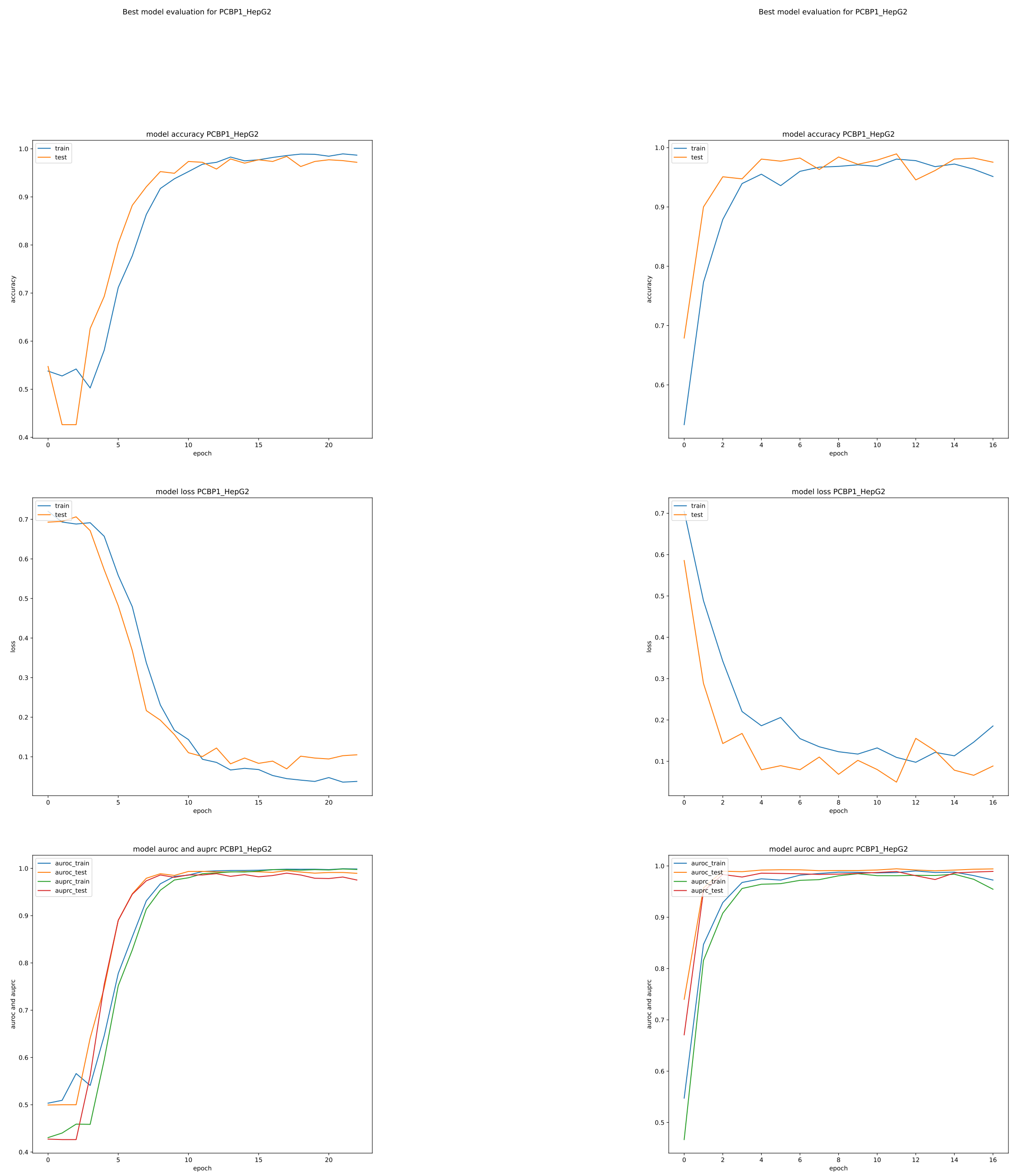


(a) Baseline Model

(b) Integrated Model

Figure A.7.: Comparison of Baseline and Integrated Model: Training

Comparison of performance metrics during training for the baseline and integrated models: Accuracy (top), Loss (middle), and Area Under the Receiver Operating Characteristic Curve (AUROC) and Precision-Recall (PR) values (bottom) plotted over the training epochs.



(a) Baseline Model

(b) Integrated Model

Figure A.8.: Comparison of Baseline and Integrated Model: Training

Comparison of performance metrics during training for the baseline and integrated models: Accuracy (top), Loss (middle), and Area Under the Receiver Operating Characteristic Curve (AUROC) and Precision-Recall (PR) values (bottom) plotted over the training epochs.

A.2.4. Setting B

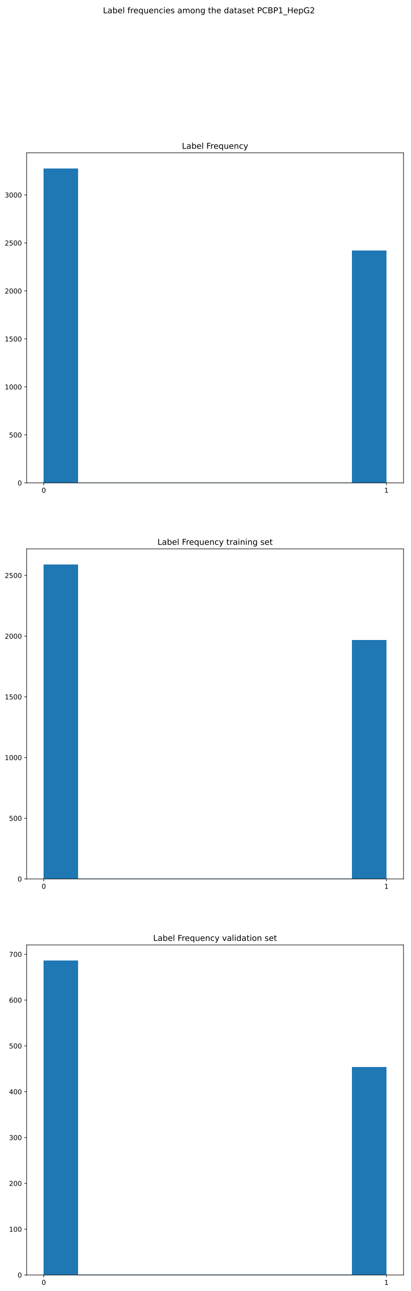


Figure A.9.: Label frequency for Setting B Augmented dataset Negative 1

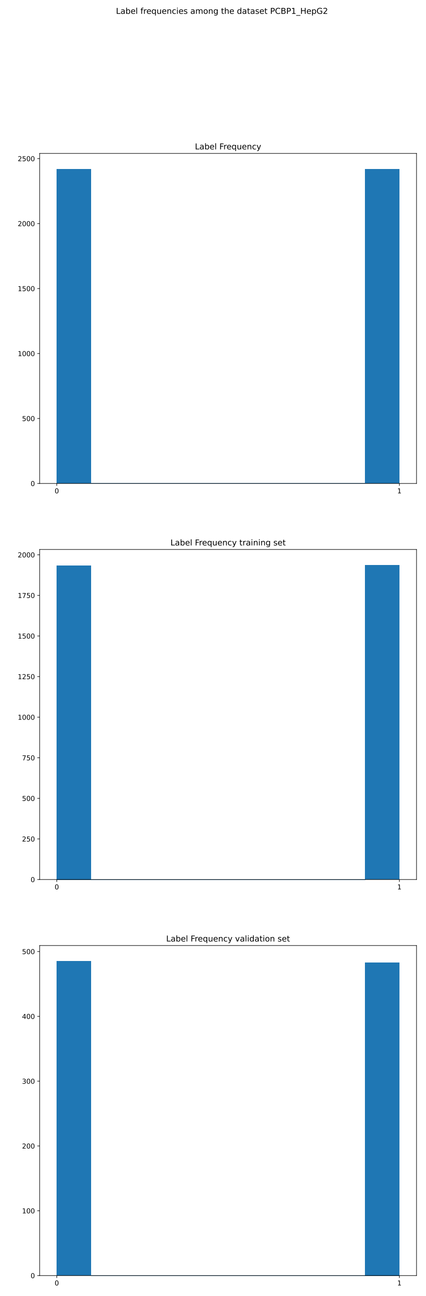
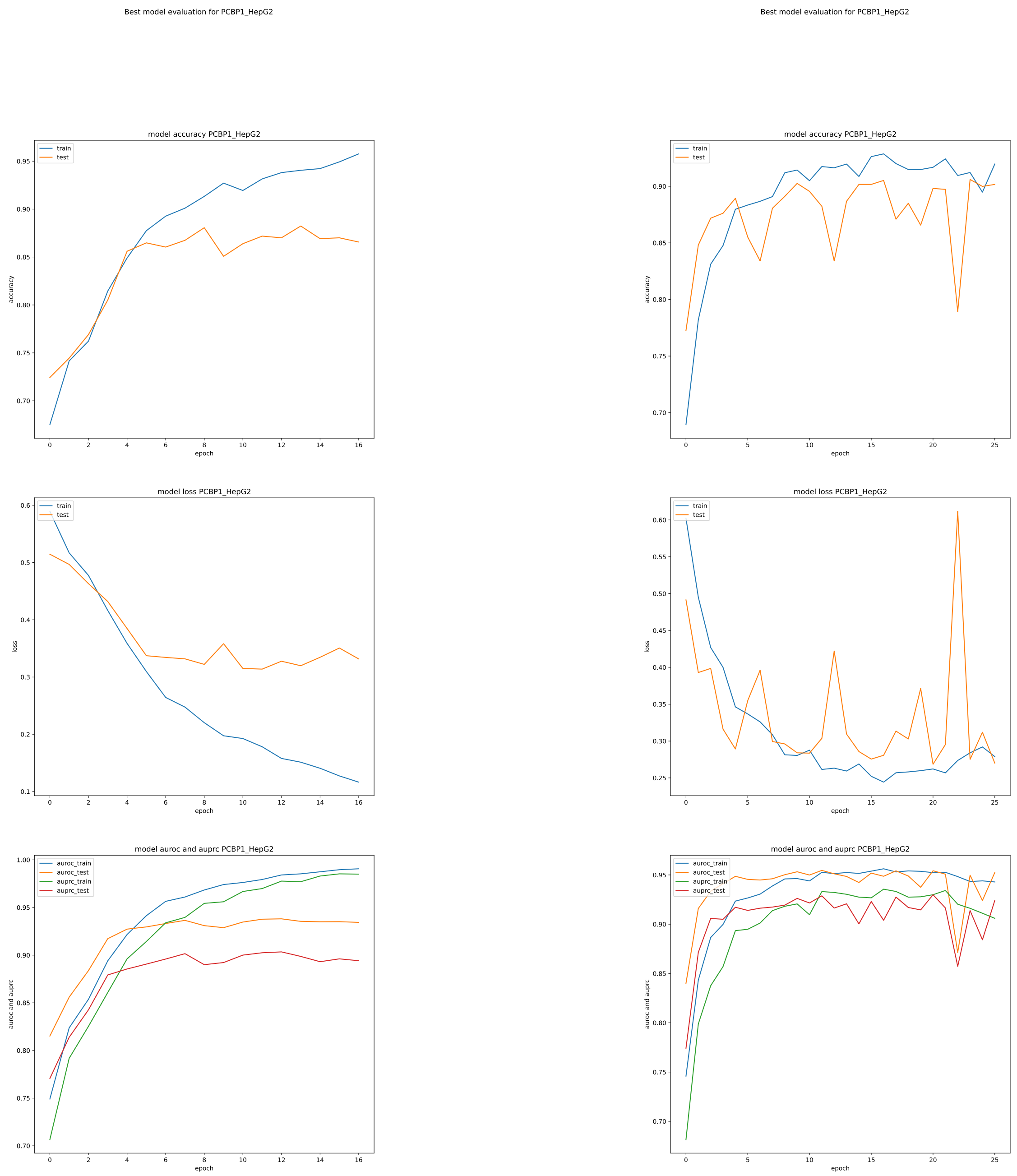


Figure A.10.: Label frequency for Setting B Augmented dataset Negative 2

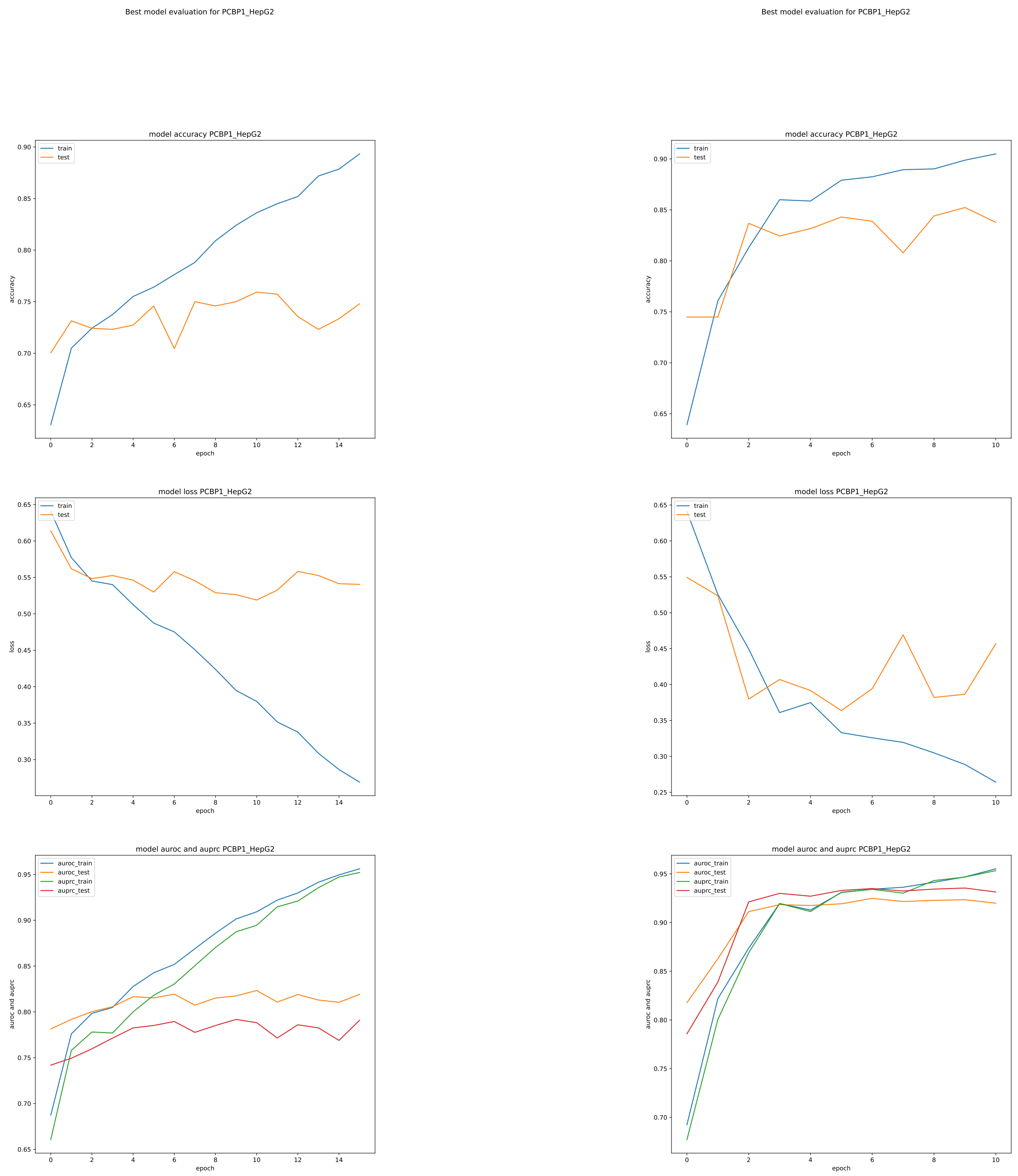


(a) Baseline Model

(b) Integrated Model

Figure A.11.: Comparison of Baseline and Integrated Model: Training

Comparison of performance metrics during training for the baseline and integrated models: Accuracy (top), Loss (middle), and Area Under the Receiver Operating Characteristic Curve (AUROC) and Precision-Recall (PR) values (bottom) plotted over the training epochs.



(a) Baseline Model

(b) Integrated Model

Figure A.12.: Comparison of Baseline and Integrated Model: Training

Comparison of performance metrics during training for the baseline and integrated models: Accuracy (top), Loss (middle), and Area Under the Receiver Operating Characteristic Curve (AUROC) and Precision-Recall (PR) values (bottom) plotted over the training epochs.

A.3. Discussion

Setting	A neg1	A neg2	B neg1	B neg2
	AKAP1_HepG2	AKAP1_HepG2		
	LARP4_HepG2		UPF1_HepG2	
	PCBP1_HepG2	PCBP1_HepG2	PCBP1_HepG2	
	METAP2_K562	METAP2_K562	CPEB4_K562	
		LARP4_K562	LARP4_K562	
	PCBP1_K562		PCBP1_K562	
	RBM15_K562		ZC3H11A_K562	

Table A.4.: Protein sets with Improved Performance in the Integrated Model compared to the Baseline Model