Saarland University

Master Thesis

Unravelling DNA Base Composition Behind Budding Yeast DNA Replication Origins Using Large Language Models

Zohreh Piroozeh

supervised by

Dr. Alina Bazarova

co-supervised by

Univ.-Prof. Dr. Olga V. Kalinina

February 26, 2025

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken,_____ _____

(Datum/Date) (Unterschrift/Signature)

Abstract

In living cells, DNA replication begins at multiple genomic sites called replication origins. Identifying these origins and their underlying base sequence composition is crucial for understanding the replication process. Existing machine learning methods for origin prediction often require labor-intensive feature engineering or lack interpretability.

In this study, we employ genome-based pre-trained LLMs to predict replication origins in budding yeast. By leveraging pre-training, LLMs automatically capture complex genomic patterns, eliminating the need for extensive feature engineering. The attention mechanism further enables the recognition of important sequence dependencies and patterns.

We fine-tuned the pre-trained DNABERT and DNABERT-2 models for our downstream task. To reveal the DNA base composition behind replication origins, we emphasize data engineering and explainability, rather than solely using models for prediction. Therefore, we evaluate model performance across datasets of varying complexity using a structured data engineering strategy, to ensure robustness.

We developed a comprehensive pipeline for identifying sequence motifs using attention maps and bioinformatics post-processing, making DNABERT more interpretable. We also discussed explainability of the attention maps extracted by DNABERT-2, as well as its learning mechanism using various approaches.

Acknowledgments

I would like to express my sincere gratitude to my supervisor, Dr. Alina Bazarova, for her invaluable support and guidance throughout this project. From the very beginning, there were moments when I felt overwhelmed and stuck, but her unwavering encouragement and willingness to help made all the difference. She was always approachable, and supportive, creating a welcoming environment for my questions and concerns.

Over the past year, I have learned a great deal from her, not only in terms of technical research skills but also in essential soft skills that have helped me grow academically. Her patient assistance during the writing process of our paper has been invaluable, and I have learned a great deal from her about academic writing. I am truly grateful for the opportunity to learn under her guidance.

I am particularly grateful to Prof. Dr. Olga Kalinina for her invaluable support and guidance as my co-supervisor throughout this project. Her insightful ideas and valuable feedback greatly contributed to improving our research results.

This work was supported and funded by Jülich Research Center, Institute for Advanced Simulation (IAS). I would like to acknowledge the valuable computing resources and support provided, which significantly contributed to the success of this research.

Finally, I also wish to extend my heartfelt thanks to my family and my spouse, Amir, whose unwavering support has been a constant source of strength and encouragement throughout this journey.

Contents

Chapter 1

Introduction

DNA replication is the fundamental biological process by which a single DNA molecule is duplicated to generate two identical copies. In eukaryotic cells, this process begins at multiple genomic sites called replication origins (ORIs), which are distributed throughout the genome. Accurately identifying these origins and analyzing their DNA base composition is essential for gaining deeper insights into the molecular mechanisms governing DNA replication.

In this chapter, the main concepts of DNA replication process are reviewed with more concentration on replication origins of saccharomyces cerevisiae (budding yeast). Then a brief introduction of genome language models is provided, following with our motivations for applying these models to predict ORIs locations in yeast.

1.1 Biological Background of DNA Replication

DNA replication is a process essential for all living organisms, ensuring the transmission of genetic information during cell division and across generations. It occurs during the S-phase of the cell cycle, preceding cell division. Although DNA sequences are replicated with high fidelity, occasional mutations can still occur.

The DNA double helix is typically stable, with its two strands held together by numerous hydrogen bonds between complementary bases. For replication to proceed, the helix must be unwound, and the strands separated to expose the unpaired bases. This process is initiated by the enzyme helicase, which breaks the hydrogen bonds between bases, creating replication forks that extend bidirectionally from the replication origin.

Several proteins work together at the replication fork to facilitate DNA synthesis. The primary enzyme, DNA polymerase, constructs new DNA strands by adding nucleotides complementary to each template strand. This self-correcting enzyme catalyzes nucleotide polymerization in the 5'-to-3' direction, ensuring accurate replication. Due to the antiparallel nature of DNA, replication proceeds differently on the two strands. On the leading strand, DNA polymerase synthesizes continuously in the 5'-to-3' direction. However, on the lagging strand, replication occurs in short, discontinuous segments using a "backstitching" mechanism, forming Okazaki fragments. Since DNA polymerase cannot

initiate synthesis on its own, RNA primers are required to start each fragment. These primers are later removed and replaced with DNA.

In summary, DNA replication involves the coordinated action of multiple proteins, each playing a crucial role in the process:

- DNA polymerase and primase: responsible for catalyzing nucleotide addition and initiating synthesis with RNA primers.
- Helicases and single-stranded DNA-binding (SSB) proteins: helicases unwind the DNA helix, while SSB proteins stabilize the exposed single strands to prevent reannealing.
- DNA ligase and primer-removing enzymes: DNA ligase joins Okazaki fragments on the lagging strand, while other enzymes remove RNA primers and replace them with DNA.
- Topoisomerases: prevent supercoiling and relieve tension in the DNA strand by resolving tangling and overwinding.

Together, these proteins form a highly coordinated "replication machine" at the replication fork, ensuring efficient and synchronized DNA synthesis[1].

1.2 DNA Replication Origins

As just pointed out, DNA replication origins are scattered across multiple locations along the chromosomes of eukaryotic cells. This enables replication to initiate simultaneously at various sites, speeding up the process. In contrast, prokaryotes typically have a single origin of replication.

Many organisms utilize specific regions in the genome as starting points for replication, indicating that the precise positioning of these origins is biologically significant. The regulation of replication origin locations is essential for maintaining coordination between DNA replication and other cellular processes. Proper coordination is important to prevent issues like DNA strand breaks and other types of DNA damage [2].

In simple cells like bacteria or yeast, replication origins are primarily determined by specific DNA sequences, which are easy to open with ability of attracting initiator proteins. A/T-rich regions are commonly found at replication origins because adenine (A) and thymine (T) pairs are held together by only two hydrogen bonds, making them easier to separate compared to G/C pairs, which have three hydrogen bonds. This lower bond strength facilitates the unwinding of DNA, allowing replication to initiate more efficiently.

The replicon hypothesis, proposed six decades ago, describes how chromosomal DNA synthesis is regulated in E. coli [3]. The model suggests that a trans-acting initiator protein interacts with a cis-acting replicator to start DNA replication at a nearby origin of replication (Figure 1.1, A, i). With the help of other proteins, such as co-loaders, the initiator deposits replicative helicases onto the DNA. Then helicases help recruit other parts of the replisome (the machinery for DNA replication), leading to the complete assembly of the replication apparatus (Figure 1.1, A, ii). The replicator determines where replication begins, while the section of the chromosome replicated from a single origin is

called a replicon. The placement and arrangement of replication origins across the genome define replicons.

A key feature of the replicon hypothesis is that DNA replication is controlled by positive regulation, meaning the replication process is actively initiated rather than being suppressed until allowed. Though the replicon hypothesis emphasized positive regulation, later research has shown that both positive and negative regulatory mechanisms control DNA replication, adding layers of complexity. These mechanisms ensure that DNA replication is tightly controlled in both bacteria and eukaryotes, in terms of timing and location within the genome [2]. The idea of the replicator as a genetic unit has been highly valuable in efforts to pinpoint replicator DNA sequences and initiator proteins in prokaryotes. It has also been useful in eukaryotes, though the structure and complexity of replicators vary significantly between these two domains of life.

In bacteria, replication typically starts from a single region, known as the replicator, which is clearly defined by a specific DNA sequence. This consensus DNA sequence element directly controls the replication process in bacterial chromosome (Figure 1.1, A). In contrast, in most eukaryotic organisms (except for budding yeast), replicators are not defined by a specific consensus sequence. Instead, these regions are determined through a combination of factors, such as the local structure of the DNA and chromatin signals. This means that replication origins in eukaryotes are more flexible and influenced by multiple elements rather than a fixed DNA sequence alone [2]

Eukaryotic chromosomes are significantly larger than bacterial chromosomes, which necessitates starting DNA replication from multiple origins at the same time to ensure the entire genome is copied timely. In addition, more replicative helicases are loaded than those activated to initiate replication during each cell cycle (Figure 1.1, B). Therefore, eukaryotic origin activation occurs on two levels: origin licensing and origin firing. Origin licensing marks potential replication origins, while origin firing selects a subset to initiate DNA synthesis by assembling the replication machinery. Extra licensed origins act as backups, activating only when nearby replication forks slow down or stall, ensuring complete DNA replication under stress. This system, along with strict cell cycle control, prevents both under-replication and over-replication, maintaining genome integrity [2].

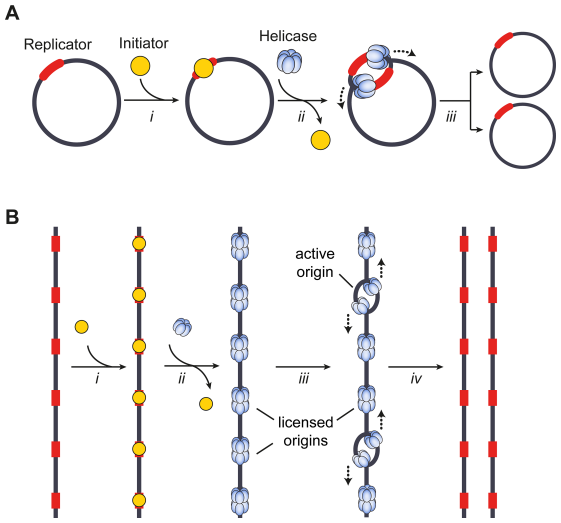


Figure 1.1: Models for bacterial (A) and eukaryotic (B) DNA replication initiation, figure from [2]

The context-based definition of replicators and origin selection in eukaryotic systems suggests a flexible replicon model, allowing variation in the DNA replication process. Although replicators and origins may be physically separated on chromosomes, they often appear close together, leading to their joint reference as "origins".

In summary, the organization, specification, and activation of origins in eukaryotes, including yeast, are more complex than in prokaryotes. Eukaryotic cells, due to their large genome sizes, require hundreds to tens of thousands of origins to initiate replication and complete DNA synthesis during each cell cycle. Additionally, in most eukaryotic cells, like humans, the location of origins is influenced by contextual cues, and there are no known consensus sequences. While *S. cerevisiae* has hundreds of origins with known locations and consensus sequences available at replication origins [2]. Moreover, DNA replication initiation are highly conserved across eukaryotes [4], suggesting that the insights gained from *S. cerevisiae* could be potentially applied to higher eukaryotes. Therefore, we focused on *S. cerevisiae* DNA replication origins to unraveling their DNA base composition using genome language models, and to evaluate models capabilities in identifying consensus sequences.

1.3 Replication origins in S. cerevisiae Genome

The features of eukaryotic replication origins are most thoroughly studied in the budding yeast (*S. cerevisiae*). In yeast, certain DNA sequences, known as autonomously replicating sequences (ARSs), enable independent replication on circular plasmids and function as replication origins (ORIs) on its chromosomes.

ARS regions are about 100–200 base pairs in length and consists of several key elements: A, B1, B2, and in some cases B3 elements, as illustrated in Figure 1.2. The A element includes the conserved 11 base pair, known as ARS consensus sequence (ACS), which, together with the B1 element, forms the main

binding site for the origin recognition complex (ORC). The B2 region has a sequence resembling the ACS and is thought to potentially serve as a secondary ORC binding site in some cases, or as a site for binding the core of the replicative helicase. On the other hand, the B3 element attracts the transcription factor Abf1, though B3 is not present at all budding yeast origins, and Abf1 binding is not considered strictly necessary for origin activity [5].

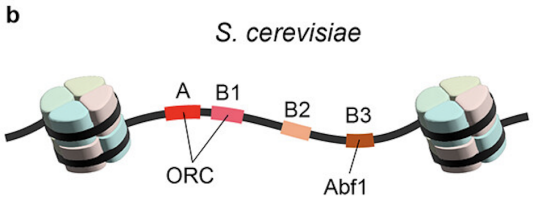


Figure 1.2: Specific replication origin sequence elements for *S. cerevisiae* [5]

The ACS was initially identified by Broach et al. (1983) through the alignment of known essential elements, defining it as an 11-bp motif represented as 5'-WTTTATRTTTW-3' [6]. This motif was later refined to 5'-WTTTAYRTTTW-3' [7]. In 1997, an extended 17-bp motif was proposed based on a larger set of origins [8]. All proposed ACS motif representations are summarized in Table 1. Here, 'W' represents A or T, 'Y' represents C or T, and 'R' represents G or A.

ACS motifs	Proposed by:
5'-WTTTATRTTTW-3'	Broach et al., 1983 [6]
5'-WTTTAYRTTTW-3'	Marahrens and Stillman, 1992 [7]
5'-WWW-WTTTAYRTTTW-GTT-3'	Theis and Newlon, 1997 [8]

Table 1.1: Known ACS motifs scheme

Existing sequence-specific DNA-initiator interactions indicate a unique mode of origin recognition in budding yeast rather than a universal method for origin specification across all eukaryotes. Although in budding yeast ORC binds to the ACS, The ACS alone is not sufficient to determine origin function, as not all experimentally identified origins contain an ACS, though the majority do. Moreover, out of 12000 ACS matches across the yeast genome, around 500 correspond to the replication origins [9]. Hence, ACS presence is neither sufficient nor necessary for the origin location, but it is the main known characteristic element of ARS in the yeast.

In recent years, the accumulation of published *S. cerevisiae* whole genome sequences has been unprecedented. Key findings from these studies can be referenced to support the explanation of results provided by genome language models, as they contribute valuable context and comparative data.

One notable study conducted a comprehensive genome-wide analysis of ORI features and provided insights into the similarities and differences of replication origin sequences across diverse budding yeast strains from a population genomics perspective [4]. In this study an analysis was conducted on the ARSs in *S. cerevisiae* S288C reference genome which illustrated 94.32% of experimentally validated ARSs were unique across the genome, with those exhibiting high

sequence similarity typically found in subtelomeric regions. A non-redundant dataset comprising 520 ARSs was established, drawing on annotations from the SGD [10], as well as data from OriDB [9] and DeOri [11] databases. A large-scale comparison of ORIs among various budding yeast strains from a population genomics perspective revealed that 82.7% of these 520 ARSs were conserved not only in sequence but also in chromosomal location, and non-conserved ARSs were primarily located in subtelomeric regions.

To better comprehend the characteristics of ARS, we can also review the comprehensive statistical analysis provided by [4]. The *S. cerevisiae* S288C reference genome includes 352 ARSs as documented in the SGD database. These experimentally confirmed ARSs have lengths ranging from 51 to 1324 bp, with the majority (63%) being between 70–250 bp. Additionally, the median length of ARSs on each chromosome is typically around 240 bp. The genome sequence of *S. cerevisiae* S288C has an average GC content of around 38%, whereas the ORI sequences exhibit a lower GC content of almost 29%. As experimental data has increased, the YeastMine database, provided by SGD, now contains 196 documented ACS sequences. Notably, approximately 87% of ARSs that include the ACS element are shorter than 300 bp. They used the ACS element as a focal point to analyze the nucleotide distribution within ARSs and visualized its base content by WebLogo as illustrated in Figure 1.3. This WebLogo reveals that the ACS is marked by a prominent central peak, characterized by a high proportion of T residues. In Addition, the upstream region exhibits a higher frequency of A residues, while the downstream region shows a predominance of T residues.

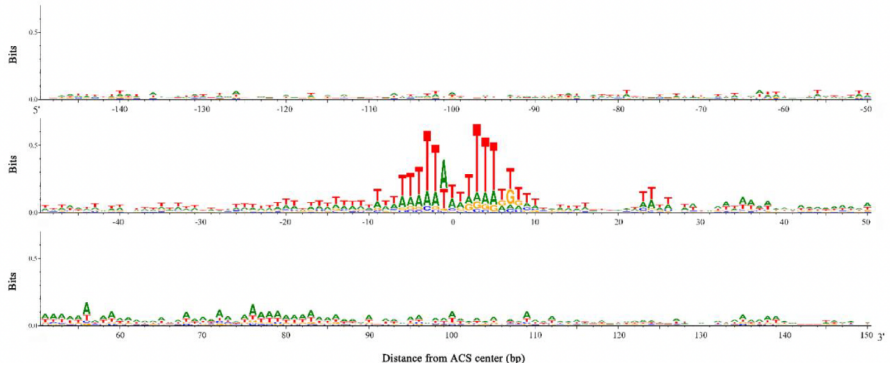


Figure 1.3: WebLogo of ARSs containing the ACS element. The 196 published ACS sequences were downloaded from the YeastMine database populated by SGD, from supplementary material of [4]

According to [4], replication origins in the *S. cerevisiae* reference genome were classified based on their positional relationships with adjacent genes. Origins located between protein-coding genes, without any intersection, were defined as intergenic ORIs, while those partially or fully overlapping protein-coding genes were classified as intersected ORIs. The results indicated that intergenic ORIs comprise 68.18% of the known replication origins in *S. cerevisiae* S288C. Analysis of the distance distribution between ARSs and their adjacent protein-coding genes revealed that most intervals are less than 1000 bp. They Identified

repeats in 92.90% of ARSs, in both intergenic and intersected ORIs. Intergenic ORIs are typically enriched with repeats characterized by continuous A bases, continuous T bases, or alternating A and T sequences (Figure 1.4), with an average AT content of 91.10% . In contrast, in case of intersected ORIs, overlapping segments containing the ACS motif within intersected ORIs and their overlapping protein-coding genes show repeats with a higher GC content, averaging an AT content of 83.52%.

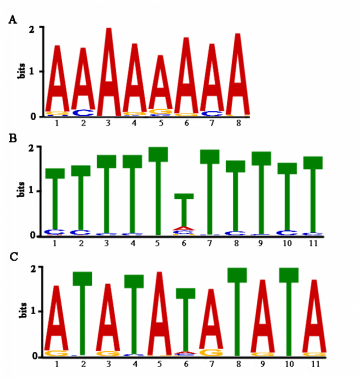


Figure 1.4: Motifs of repeats in intergenic ORIs of *S. cerevisiae* S288C, from supplementary material of [4]

1.4 Motivations to Apply Language Models

Large language models (LLMs) are a class of advanced deep learning models that utilize extensive pre-training on vast datasets to identify and learn complex patterns within data. Originally developed for natural language processing tasks, LLMs have demonstrated remarkable versatility in analysing diverse types of sequential data, including genomic sequences.

In the context of DNA replication origin prediction, current machine learning (ML) methods often rely on manually extracted features, which can be labour-intensive and may not fully capture the intricate patterns inherent in genomic data. The application of LLMs presents a significant opportunity to overcome these limitations. LLMs enhance predictive accuracy by leveraging their pre-training capabilities to automatically capture complex genomic patterns. This eliminates the need for intricate feature engineering and reduces the burden of labour-intensive feature extraction, streamlining the overall prediction process.

Recently, genome based LLMs have gained significant traction for fine-tuning pre-trained models on various downstream genomic tasks, such as gene expression prediction, enhancer classification, and mutation impact analysis. By adapting LLMs to specific genomic challenges, researchers have achieved breakthroughs in unravelling complex biological phenomena. Their ability to generalize and adapt across downstream tasks demonstrates their potential for decoding the vast complexity of genomic data. This widespread adoption motivated us to leverage genome language models for DNA replication origin prediction.

To date there are several available LLM models pretrained on genome data which have been successful in e.g., a problem of identifying Transcription Factor

binding sites and underlying DNA sequence motifs, which is conceptually similar to the identifying ACS for DNA origins. Models like DNABERT [12] and DNABERT-2 [13] were specifically developed for working with DNA sequence data and pre-trained on the human genome data and multispecies genome data, respectively.

In this study we are aiming to understand how to predict locations of origins in budding yeast, or more precisely, finetuning mentioned genome language models to distinguish between sequences including ORIs and sequences not including. In addition to comparing performance of these models, we are interested in interpretation of results to unravel DNA base composition of DNA replication origins. As mentioned before, although ACS presence is neither sufficient nor necessary condition for the origin location, it is still the main known character of ARS in the yeast ORIs. So, the key questions here are: To what extends model's performance are depended to ACS? Is there anything on top of ACS which determines the replication origin recognized by LLMs? To address these questions, we employ a data engineering strategy to evaluate the model's performance across datasets of varying complexity. Additionally, to enhance model explainability, we develop a comprehensive pipeline for identifying underlying sequence motifs using attention maps, while also incorporating attention-based and perturbation-based explainability techniques.

Chapter 2

Related works

Understanding the DNA replication mechanism can offer valuable insights into the regulation of cell division and the cell cycle. Additionally, it plays a crucial role in the development of novel treatments for various diseases. Hence, accurately identifying Origin of Replication Initiation sites (ORIs) is a fundamental step in advancing the study of DNA replication processes. While techniques like Chromatin immunoprecipitation (ChIP-seq) and next-generation sequencing technology are widely used to identify ORIs, they are costly and time-consuming, making it difficult to map ORIs across entire genomes. In this context, computational and AI-based approaches have emerged as powerful tools to complement experimental methods, offering faster, cost-effective, and scalable solutions for accurate genome-wide ORI prediction.

This chapter reviews recent advancements in computational approaches for predicting eukaryotic ORIs, with a particular emphasis on studies involving the budding yeast genome. It highlights machine learning techniques that focus on feature extraction and selection, and the predictive results achieved by these methods. While large language models (LLMs) excel at leveraging extensive pre-training to effectively identify complex genomic patterns without the need for detailed feature engineering, reviewing methods that incorporate feature selection provides valuable insights into the critical features of ORIs.

2.1 Characterizing ORIs for ML Approaches

Since eukaryotic cells have long, linear chromosomes, require multiple origins of replication to efficiently replicate their DNA during the S phase of the cell cycle. The budding yeast, is a well-studied model organism that has provided significant insights into the mechanisms of eukaryotic DNA replication. Discovery of the intrinsic characteristics is valuable for precisely identifying ORIs, since most computational methods rely on known characteristics to predict origins.

The recognition of ACS sites by origin recognition complex (ORC) proteins during replication initiation is influenced by a combination of DNA structure, chromatin state, and nucleotide composition. These DNA properties vary across the genome due to biological factors. ORC proteins recognize ACS sequences through direct interactions with nucleotide bases and indirect recognition of structural properties. Therefore, it has been suggested that both sequence com-

position and conformational properties are crucial for understanding replication mechanisms [2]. However, most of the research conducted on predicting ARSs in *S. cerevisiae* has primarily focused on sequence composition. In 2014, a sequence analysis study on yeast ORIs was conducted, uncovering several key features of ORIs [14]. They found that the GC content surrounding ORIs in the *S. cerevisiae* genome is lower than the genome-wide average, with significantly reduced GC profile and GC skew scores within ORI regions compared to their flanking regions, suggesting similarities to bacterial replication mechanisms. ORI sequences were shown to exhibit strong short-range base correlations, indicating a highly ordered structure and distinct sequence organization. ORIs were predominantly located in nucleosome-free regions. Moreover, sometimes sharing elements with promoters, or mostly ORIs are in transcription terminal regions. These findings indicate that most ORIs are not preferentially located near transcription start regions, which may help ensure the coordination between replication and transcription. Finally, using an SVM model, they demonstrated that nucleosome occupancy is a much stronger predictor of ORIs compared to GC skew, highlighting its critical role in replication initiation.

Breier et al. proposed an algorithm, Oriscan, to identify potential DNA replication origins in budding yeast [15] by comparing them to 26 well-characterized origins. Oriscan considered both the ACS motifs and the surrounding AT-rich regions. Through a ranking process that prioritizes predictions based on their similarity to known origins, Oriscan achieved an 84% matching rate with known ARSs among its top 100 predictions. However, this rate dropped to 56% when considering the top 350 predictions. These results indicate that relying solely on similarity to the 26 featured ORIs may limit the algorithm's ability to comprehensively identify ARSs.

2.2 Machine Learning Methods

Machine learning approaches for predicting ORIs in yeast genomes have gained traction in last decade. Since the physicochemical properties of oligonucleotides play a crucial role in regulating DNA replication, Chen et al. 2012 ainvestigated two characteristics surrounding replication origins in the *S. cerevisiae* genome, including DNA bendability and cleavage intensity [16]. They found that these properties significantly reduced in core replication regions compared to the upstream and downstream regions of ORIs. Leveraging these two structural DNA features, they developed a support vector machine (SVM) model to predict ORIs and tested its performance using a benchmark dataset. The results demonstrated accuracy of 85.86% through jackknife cross-validation, which demonstrated DNA bendability and cleavage intensity effectively characterize core replication regions.

In 2015, iORI-PseKNC [17] was introduced, a novel predictor based on the pseudo k-tuple nucleotide composition (PseKNC) [18], to encode DNA sequences of *S. cerevisiae* genome. The method captures not only the composition of nucleotides but also the sequence-order information and physicochemical properties of DNA. iORI-PseKNC represents DNA sequences as a vector of $4^k + \lambda$ components, which is a consist of frequencies of k-tuple nucleotide composition, and the sequence-order correlation factors, as follow:

$$\text{PseKNC} = [f_1, f_2, \cdots, f_{4^k}, \theta_1, \theta_2, \cdots, \theta_\lambda] \tag{2.1}$$

Where:

- $f_1, f_2, \ldots, f_{4^k}$: Frequencies of k-tuple nucleotides.
- θ_λ: Sequence-order correlation factors that encode the physicochemical properties across distances λ.

The parameter k captures local (short-range) sequence order effects, while λ reflects global (long-range) sequence order effects. For a sequence S, the λ-tuple correlation factor is calculated as:

$$\theta_\lambda = \frac{1}{L-\lambda} \sum_{i=1}^{L-\lambda} \Psi(i, i+\lambda) \tag{2.2}$$

Here L denotes the length of the DNA sequence, and $\Psi(i, i+\lambda)$ quantifies the relationship between two nucleotides at positions i and $i+\lambda$, using aggregation of physicochemical properties, including 6 structural properties: twist, tilt, roll, shift, slide, and rise.

iORI-PseKNC incorporates the calculated PseKNC vectors into a support vector machine for predicting ORIs in the *S. cerevisiae* genome. The predictive model accuracy is 83.72%, when the parameters k and λ are equal to 3 and 50, respectively. The method iORI-PseKNC outperforms Chen et al. 2012 [16], which only considers properties like DNA bendability and cleavage intensity.

Then in 2016, Xiao et al. introduced iROS-gPseKNC [19] by incorporating position-specific propensity into sequence representation. The approach begins by identifying all dinucleotides (e.g., AA, AC, AG, etc.), totaling 16 combinations. For DNA sequences with length 300, a matrix of dimensions 16× (300-1) is constructed. $P_i,_j$ (each element of matrix) represents the difference in the occurrence frequency of the i-th dinucleotide at the j-th position between two datasets, positive and negative datasets. The values in the matrix quantify the propensity of each dinucleotide at specific positions in sequences with replication origins compared to those without origins. This matrix is then used to represent each DNA sequence in a vector into general form of PseKNC, as shown in 2.3, when u-th component of this vector is selected as 2.4

$$\mathbf{D} = \begin{bmatrix} \phi_1 & \phi_2 & \cdots & \phi_u & \cdots & \phi_z \end{bmatrix}^T \tag{2.3}$$

$$\phi_u = \begin{cases} P_{1,u} & \text{when } N_u N_{u+1} = \text{AA} \\ P_{2,u} & \text{when } N_u N_{u+1} = \text{AC} \\ P_{3,u} & \text{when } N_u N_{u+1} = \text{AG} \quad (1 \le u \le 299) \\ \vdots & \vdots \\ P_{16,u} & \text{when } N_u N_{u+1} = \text{TT} \end{cases} \tag{2.4}$$

Then Random Forest classifier is used, achieving significant accuracy of 98.03%. The predictor outperforms iORI-PseKNC method on the same dataset (sequences with length 300 bps), highlighting the importance of incorporating position-specific information into sequence analysis for improved prediction of biological sites.

While existing predictors of origins of replication (ORIs) were limited to species-specific applications and could only analyze short DNA sequences of

fixed lengths (typically 250–300 base pairs), a new method has been developed to handle diverse yeast species and accommodate sequences of varying lengths. The method iRO-3wPseKNC [20] proposed by Liu and colleagues in 2018, by encoding DNA sequences into vectors of features while incorporating GC asymmetry bias as a predictive feature and utilizing a three non-overlapping local windows approach. DNA sequences are represented in the form of PseKNC vector, which represents the normalized frequency of specific k-tuple nucleotides occurring within the front, middle, and rear windows of the sequence. Using the random forest classifier, iRO-3wPseKNC demonstrates improvements across multiple yeast species, however its performance for *S. cerevisiae* specifically does not show improvement over iORI-PseKNC.

Nevertheless, iRO-3wPseKNC was the only existing predictor that can predict the entire replication origins with significant variation in length . Then Liu et al. tried to improve results of their method by introducing iRO-PsekGCC predictor [21], which builds upon the foundation of iRO-3wPseKNC. The three-window strategy has two main limitations: (1) it captures only local GC asymmetry bias, missing global GC asymmetry patterns, and (2) using large k-values in k-tuple nucleotide analysis results in high-dimensional feature vectors, leading to computational challenges. To address these issues, the study introduces a new feature representation called k-tuple GC composition (k-GCC), which effectively captures GC preferences in replication origins along with their global interactions. This enhancement enables the model to better utilize the "GC asymmetry bias" and sequence order effects, resulting in more accurate identification of DNA replication origins.

In 2019, another method, developed as iORI-PseKNC2 [22], enhanced prediction accuracy by improving PseKNC (type I) [18] and a two-step feature selection. Unlike PseKNC, which uses a fixed number of components $(4^k + \lambda)$, Type II PseKNC expands the feature vector to $4^k + n\lambda$, where n represents the number of physicochemical properties being considered, incorporating them more comprehensively into the feature representation. First, by considering 90 physicochemical properties encoded into PseKNC, to characterize DNA sequences. Then F-score [23] and mRMR [24] were utilized to optimize feature selection in sequence representation to eliminate redundant and noisy data. This approach aimed to include higher order correlations between features while minimizing redundancy. Like iORI-PseKNC, SVM was employed for classification, improving accuracy to 87.79%. This finding underscores the effectiveness of two-step feature selection in enhancing prediction accuracy and decreasing the dimensionality of the feature vector.

Singh and colleagues compared three machine learning algorithms (KNN, NB, and SVM), by proposing a multi-view ensemble learning (MEL) approach to identify DNA replication origins in yeast [25]. By leveraging multiple DNA segment properties, such as sequence composition, physical properties, and structural characteristics, they created diverse feature subsets or "views". Each view was used to train classification models, the MEL models produce predictions by computing a weighted average of multiple classifiers, each trained on different optimal feature subsets. However, they found SVM as a better choice for prediction of ARS compared to NB and KNN.

In another attempt to address the problem of ORI prediction using machine learning methods, the extreme gradient boosting (XGBoost) approach was proposed in 2020 [26]. In this study a hybrid feature set is used, combining various

biological and sequence-based properties. They compared the performance of XGBoost with three methods including: DNA bendability and cleavage intensity around ORIs [16], iORI-PseKNC [17], and iORI-PseKNC2.0 [22]. As shown in table 2.1, it can be concluded that the proposed method performs better. Although the study shows that integrating diverse feature types improves ORI prediction accuracy, like most previous works, it requires intensive feature selection.

Method	Sensitivity	Specificity	Accuracy	MCC
Bendability + cleavage	81.23	80.3	80.76	0.6153
Type-I PseKNC	84.69	82.76	83.72	0.6746
Type-II PseKNC	89.63	85.96	87.79	0.7564
XGBoost	85.19	93.83	89.51	0.7931

Table 2.1: Comparative performance among different predictors [26]

In 2021, yORIpred was proposed as a computational prediction method for DNA replication origins (ORIs) in yeast species, leveraging advanced feature representation and machine learning tools [27]. Five different classifiers (ANN, RF, GB, ERT, and SVM) and eight feature encoding methods (Kmer, CKSNAP, DPCP, PseDNC, TPCP, PseKNC, SCPseKNC, and EIIP) were utilized to evaluate their contributions to ORI prediction across yeast species. Each encoding was individually paired with the classifiers, and a 10-fold cross validation was applied to develop respective prediction models. In total, 40 prediction models (8 encodings $\times$ 5 classifiers) were generated for each species, and the predicted ORI probabilities were concatenated into a 40-dimensional (40D) feature vector. Finally, the authors used the outputs of these 40 models as input features for an iterative feature learning to develop a final prediction model. In fact, the 40D probabilistic feature vector, input to ML classifier and the probabilities predicted by this model were treated as a new feature vector, which was combined with the original 40D vector to create a 41D feature vector. This iterative process was repeated up to 11 rounds, improving prediction accuracy and creating the optimized predicted probability of ORIs. The results present a performance comparison between yORIpred and existing predictors (iRO-3wPseKNC and iRO-PseKGCC), demonstrating that yORIpred achieves superior predictive performance compared to the other methods. However, the procedure of yORIpred involves multiple layers of manual intervention, high-dimensional feature engineering, and computationally expensive iterative steps. While the approach achieves accurate predictions, the complexity and resource demand make it cumbersome compared to more streamlined, end-to-end predictive methods like deep learning.

2.3 Deep Learning Methods

While traditional machine learning approaches for ORIs prediction have advanced the field, they have limitations. They primarily focus on local DNA sequence information, neglecting long-range interactions in 3D space. Their performance is highly dependent on manually selected features, which may be insufficient. Finally, machine learning models (like SVM and Random Forest)

treat feature extraction and classification as separate processes, which may limit predictive performance and generalization. So deep learning has gained significant attention to address the problem as a more robust, automatic framework that considers global sequence information and integrates feature extraction and classification is needed.

To overcome the limitations of prior research, a deep learning method was proposed in 2021 [28]. Wu et al. introduced sequence segmentation and utilized the Word2vec word embedding technique that captures intrinsic relationships within sequences of varying lengths. A convolutional neural network (CNN) with an embedding layer is then used to construct a deep learning framework for ORI identification. This approach leveraged Word2Vec's ability to preserve inner correlations in sequence features while enabling advanced pattern recognition by CNN. To prove if feature vectors constructed can effectively characterize the inner relationship among trinucleotides in ORIs, the t-distributed stochastic neighbor embedding (t-SNE) algorithm is applied to the original feature vectors, and reduction dimensionality diagrams demonstrates Word2vec's effectiveness in converting the relationships between words into numerical features. The proposed method achieves notable performance improvements in ORI identification for four yeast species. Specifically, it achieves an accuracy of 97% for *S. cerevisiae*, outperforming the XGBoost model [26], which achieves an accuracy of 89%. While it achieves high accuracy, it requires training from scratch, making it computationally expensive, especially for large genomic datasets, since training Word2Vec embeddings itself requires significant time and computational resources. Moreover, the integration of deep learning and embeddings makes the model less interpretable. Traditional machine learning models like support vector machines (SVMs) or decision trees often allow for easier understanding of how features contribute to predictions.

In 2022, to address the problem of origin of replication (ORI) prediction, ORI-Deep was introduced as a deep neural network designed to identify ORIs across four different eukaryotic species [29]. To enhance data representation, the model constructs a feature vector using statistical moments derived from genomic data, which are then fed into a long short-term memory (LSTM) network for training and testing. Rigorous validation of the model showed that ORI-Deep outperforms traditional ML methods in accuracy. However, the model's reliance on manually crafted features, such as statistical moments, introduces potential limitations. Manual feature extraction assumes that the selected features sufficiently represent the complexity of DNA sequences, which can lead to bias. LSTMs are highly effective at learning directly from raw sequential data, as they can capture complex patterns without relying on pre-defined features. However, ORI-Deep uses pre-constructed feature vectors instead of raw DNA sequences, which may limit the model's ability to uncover hidden sequence dependencies. Moreover, manually extracted features are often tailored to specific training datasets and may not adapt well to different organisms or sequence characteristics. In contrast, models like convolutional neural networks (CNNs) or transformer-based architectures automatically learn hierarchical features directly from raw sequences, offering greater flexibility and applicability across diverse datasets.

In one of the latest studies, Yin et al. introduced Ori-FinderH, a computational approach designed to efficiently and precisely predict human ORIs of various lengths [30]. The method combines the Z-curve method, a geometri-

cal representation of DNA sequences, with a deep learning framework for enhanced ORI prediction. The inclusion of a self-attention mechanism enhances the model's ability to focus on relevant features within the input data, leading to exceptional performance in predicting ORIs of various lengths. Developed model is generalizable across multiple cell lines, and the cross-cell-line predictive model performance is significant, with an AUC of 0.97 and accuracy of 0.91. The study does not explicitly mention an analysis of the explainability of the deep learning model. However, researchers employed a genetic algorithm (GA) combined with the proposed deep learning model to develop a process for generating artificial ORIs. In fact, the genetic algorithm uses the trained deep learning model as a fitness function to evaluate the quality of generated sequences. Sequences predicted to contain ORIs are selected and evolved iteratively to improve their ORI-like properties. These artificial ORIs were evaluated using a third-party ORI prediction tool, yielding highly favorable results, and further demonstrating the model's utility and robustness.

While deep learning approaches have shown considerable performance improvements over traditional machine learning, their requirement for training from scratch makes them computationally expensive. This also negatively impacts model interpretability, a crucial factor in identifying DNA replication origin motifs.

Chapter 3

Dataset

Over the last two decades, as experimental data and sequencing genomes have grown and more sophisticated computational tools have been proposed, numerous databases have been developed to manage genetic data. Among these, some databases are specifically developed to store information related to genome replication origins. In this chapter, we provide a brief overview of the existing databases for replication origins of budding yeast and describe how we constructed our datasets based on OriDB [9].

3.1 Data Sources for S. cerevisiae ORIs

The identification of ORIs involves origin-associated proteins or DNA synthesis activity at active replication sites [9]. Protein-based identification of origins of replication (ORIs) relies on Chromatin Immunoprecipitation (ChIP) to detect key proteins like ORC, which binds to replication origins and initiates DNA replication, and Minichromosome Maintenance (MCM) proteins, essential components of the replication machinery. By isolating and sequencing DNA fragments bound to these proteins, researchers can locate ORIs. While, sequence-based methods, such as computational searches for sequence motifs like the ACS in budding yeast, identify hallmark sequences of replication origins. Comparative genomic approaches further refine these predictions by identifying conserved sequences across related species. Finally, replication timing analysis identifies active ORIs as early replicating regions during the cell cycle, and genome-wide measurements of replication timing highlight these regions as potential ORIs.

Among the various genomic datasets available, only a few specifically include information on origins of replication (ORIs) for *S. cerevisiae*, such as OriDB [9], SGD [10], DeOri [11].

OriDB is a comprehensive database designed to collate genome-wide mapping studies of DNA replication origins, including reliable and high-quality data curated from experimental studies. Initially focused on *S. cerevisiae* ORIs, then database was significantly updated and then expanded to include replication origin mapping studies for S. pombe (fission yeast), alongside *S. cerevisiae*.

SGD (Saccharomyces Genome Database) is a comprehensive resource for the yeast genome, covering annotations for genes, replication origins, regulatory elements, and pathways. While its focus is broader, it includes detailed

information on ORIs as part of the genomic annotation.

DeOri includes a wide range of experimentally validated ORIs across multiple eukaryotic species. It compiles genome-wide data from experimental studies and is designed for cross-species comparisons. It is considered a key resource for understanding replication dynamics across various eukaryotic species, including *S. cerevisiae*. However, it may not provide sufficient depth for individual species compared to focused datasets like OriDB.

Since this study focused on *S. cerevisiae*, we found OriDB to be a robust and well-documented resource, as it provides highly curated and specific data with experimental evidence, making it suitable for benchmarking and comparison of results across multiple studies.

The OriDB database integrates data from various published studies, including five that were used to collect *S. cerevisiae* data [31]. These studies employed advanced experimental techniques to identify DNA replication origins . Four of them used microarray techniques to independently map approximate origin locations across the yeast genome. These techniques include Origin Recognition Complex (ORC) Binding Assays, DNA Replication Timing Profiles, and Two-Dimensional Gel Electrophoresis. The fifth study employed phylogenetic conservation analysis to identify a separate set of origin sites. OriDB then integrated the data from all five studies to create a consolidated list of origin sites, assigning each proposed site a status (confirmed, likely, and dubious) reflecting the confidence that it represents a true origin [31].

3.2 Datasets Based on OriDB

To construct datasets, we utilized the OriDB database, which contains 829 replication origins categorized into three groups: confirmed, likely, and dubious. To ensure a reliable dataset, we focused solely on the confirmed origins. Then sequences shorter than 500 base pairs (bp) are selected, providing us 325 confirmed ORIs. This selection aligns with the input size limitation of the DNABERT model. For sequences shorter than 500 bp, we extended them unsymmetrically and randomly on both the left and right sides to reach a uniform length of 500 bp. These sequences were labeled as 1, representing origin-including sequences or positive samples. Given our use of attention-based explainability, extending shorter sequences to 500 bp allows to investigate the distribution of attention scores. This approach enables us to assess the relative importance assigned by the model to sequence features within the origin region compared to the extended flanking regions.

Like most genome analysis studies, replication origins constitute a small portion of the entire genome, making the generation of non-origin samples challenging because it directly impacts model performance and its ability to differentiate origins from non-origins. In related studies, various strategies are employed to create non-origins or negative samples. For instance, in some studies, like [22] and [26], negative samples are generated randomly from non-replicative regions of the genome, ensuring no overlap with annotated origins While some researchers prefer to focus on immediately flanking regions of ORIs, which are close to but not within replication origins [25].

Each mentioned approach for constructing non-origin samples of dataset has its strengths and limitations, which can influence the performance and reliability

of the predictor models. These two approaches can be discussed as follow:

Randomly Subsampled Negatives: The distinction between positive (origins) and negative (non-origins) samples is much clearer. Additionally, the diversity of the negative samples ensures the dataset captures a wide range of non-origin features, reducing potential bias toward specific non-origin sequences. While this approach facilitates straightforward classification, the resulting predictor model may not generalize well to more challenging scenarios. Specifically, it might fail to differentiate origins from sequences with subtle similarities, as these might be underrepresented in a random process. Consequently, the model could lack robustness when applied to more ambiguous cases.

Negatives Resembling Origins (e.g., flanking regions or motif-similar sequences): Including negatives that closely resemble origins forces the model to identify subtle differences, testing its capacity for sophisticated discrimination. This approach enhances robustness, as it better simulates the complexities of real-world genomic analysis, where such distinctions are crucial. However, the inclusion of highly similar negatives increases the rate of false positives (misclassify non-origins as origins) and false negatives (failing to identify true origins). This can reduce the predictive accuracy of the model and complicate its interpretation, as the decision boundaries become less distinct.

Since the primary focus of this study is on explainability, the negative sample generation process has been designed to align with these objectives. To specifically challenge the model's ability to distinguish between origin and non-origin sequences, we employed a data engineering strategy to subsample diverse non-origin sequences at varying levels of complexity. This provides us with several datasets containing the same positive instances but different negative instances that vary in difficulty and properties, thus enabling robust interpretability. Each dataset was generated according to the following procedure:

Random-Neg dataset: To construct the random negative examples, we first considered the 325 confirmed OriDB origins as our positive instances. For the negative instances, we randomly subsampled 325 non-overlapping regions from the genome that do not intersect with any positive ranges. These randomly selected sequences serve as the negative samples, and by combination of positive and negative samples our Random-Neg dataset is created for distinguishing between origin and non-origin sequences.

ACS-Neg dataset: In this dataset, we selected origin and non-origin sequences in a way that include at least one ACS motif match. We queried the budding yeast genome using Homer [32] and identified 11500 matches to ACS motifs. First, to ensure all positive samples include ACS matches, we filtered out of our primary selected OriDB origins; out of 325 ORIs 298 had intersections with ACS motifs matches (or remarkably close to ACS motifs). These 298 ORIs were considered as positive samples. Then 298 negative samples were generated by selecting from the ACS motifs which are non-replicating ,and not located near any experimentally validated replicating ACS matches. These motifs were extended to 500 bp so that they do not overlap with positive samples. Hence, in this dataset, both origins and non-origin sequences include ACS motifs. The aim is to investigate whether there are additional discriminative factors, beyond ACS motifs, that can distinguish origin sequences from non-origin sequences us-

ing genome language models.

Shuffled-Neg dataset: To examine the significance of base pair order and bold patterns in distinguishing origin sequences, we developed the Shuffled-Neg dataset. Importantly, while the base pair content and frequency remained identical to the positive instances, the characteristic patterns and motifs were broken. This was achieved by shuffling the base pairs of positive samples (origin sequences) to generate negative samples. The shuffling was performed at the base-pair level prior to tokenization, ensuring that both local and global sequence patterns were disrupted. Using the 325 OriDB-confirmed origins as the starting point, we generated 325 corresponding shuffled sequences to serve as negative samples.

Block-5-Shuffled-Neg Dataset: This strategy is like the Shuffled-Neg dataset, but instead of shuffling at the base-pair level, positive samples are segmented into blocks of 5 base pairs, and these blocks are then shuffled. The objective of this approach is to preserve short-range patterns within the blocks while disrupting long-range patterns across the sequence. This allows for an investigation into the significance of short- versus long-range sequence patterns in distinguishing origin sequences. Using all the 325 OriDB-confirmed origins, 325 corresponding shuffled sequences generated as negative samples.

Table 3.1: Overview of Datasets

Dataset	**Positive (origin) samples**	**Negative (non-Origin) samples**
Random-Neg	325 confirmed origins	325 randomly selected non-overlapping origins
ACS-Neg	298 confirmed origins, with replicating ACS	298 extended non-replicating ACS
Shuffled-Neg	325 confirmed origins	325 subsampled by shuffling positive samples
Block-5-Shuffled-Neg	325 confirmed origins	325 subsampled by shuffling positive samples (blocks of 5 bp)

These four datasets can be categorized into two groups:

- Datasets with negative subsampling using genome fragments that do not replicate, including the Random-Neg and ACS-Neg datasets.
- Datasets with artificially generated negative samples, created by shuffling the origin replicating sequences.

3.3 Dataset Splitting

To finetuning and evaluate the performance of genome language models, the dataset is divided into training, validation, and test sets using two distinct strategies: random splitting and chromosome-based splitting. Both strategies provide complementary insights into the model's robustness and generalizability.

A random splitting approach is used for all types of our datasets. When sequences are randomly assigned to the training, development (dev), and test sets, ensuring that data from all genomic regions are represented across all subsets.

To enhance robustness, we used seven different random seeds, generating seven distinct splits for each dataset. The performance of the models is evaluated on all seven splits, and the average performance is reported. Each dataset was divided as follows:

- Training set: 70% of the total dataset
- Development (dev) set: 10% of the total dataset
- Test set: 20% of the total dataset

This method provides a balanced distribution of sequences and maximizes data usage but is prone to data leakage due to shared sequence similarity across subsets. Alternatively, chromosome-based splitting assigns entire chromosomes to either the training or testing sets, ensuring that sequences in the test set are entirely independent of those used for training. While this method avoids overfitting and data leakage since some chromosomes are not seen by models during training. However, it introduces the challenge of working with imbalanced data distributions and potentially less diverse training data.

Therefore, we used chromosome-based splitting for Random-Neg dataset. Seven different splits are applied, each assigning chromosomes uniquely to the training, development (dev), and test sets. In each split, a different non-overlapping combination of chromosomes was allocated to train, dev, and test, to ensure diverse and independent sets.

Chapter 4

Methods

For decades, as computational approaches for analyzing biological sequential data have developed, bioinformatics has benefited from natural language processing (NLP). Just as grammatical and semantic structures define natural languages, nucleotide composition and sequence structure determine the motifs and functions of gene sequences. LLMs, with their extensive pre-training, can effectively capture complex patterns in genomic data, enhancing predictive accuracy without the need for intricate feature engineering.

This chapter opens with an introduction to the BERT model, which serves as the foundational architecture for the language models utilized in this study, followed by an overview of DNABERT and DNABERT-2, and their architectures. Section 4.2 explains how these models were applied for our specific downstream task . Finally, we presented the explanation methodologies employed and our proposed pipeline for enhancing the explainability of models for unravelling DNA base composition behind budding yeast DNA replication origins.

4.1 Background of Methods

DNABERT and DNABERT-2 are genome language models with architecture based on the original BERT (Bidirectional Encoder Representations from Transformers) model [33], adapted for genomic data. This section provides a brief overview of the BERT model.

BERT's architecture is a multi-layered, bidirectional Transformer encoder, based on the Transformer architecture [34]. The encoder consists of multiple stacked layers, all sharing the same architecture, including multi-head self-attention mechanisms and feed-forward neural networks. Each sub-component (both the self-attention and the feed-forward network) is followed by a normalization layer. While models like GPT process text in a one direction (left-to-right or right-to-left), BERT is designed to pretrain bidirectional representations from unlabeled text, by considering both preceding and following context at every layer, enabling the pre-training of a deep bidirectional Transformer.

The BERT model used the paradigm of pre-training and fine-tuning, as illustrated in Figure 4.1 . First, the model is pre-trained on unlabeled data using pre-training tasks such as Masked Language Modeling (MLM) and Next Sentence Prediction (NSP). After the initial pre-training, the model is further

trained in a process called fine-tuning, a form of transfer learning. Then the model is initialized with the pre-trained parameters, and all parameters are then fine-tuned using labeled data specific to each downstream task. Each downstream task thus results in its own fine-tuned model, although all are initialized with the same pre-trained parameters. A key characteristic of BERT is its consistent architecture across different tasks, with minimal differences between the pre-trained and final downstream architectures.

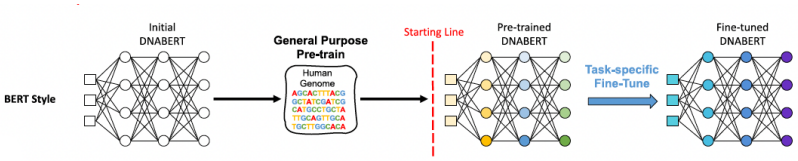


Figure 4.1: Pretraining and fine-tuning paradigm in BERT style language models, figure from [12].

In an MLM task, some words in a sentence are randomly masked (replaced with a special token like [MASK]), and the model is trained to predict these masked words based on the surrounding context. In the pre-training phase of the BERT model, alongside the MLM, a "Next Sentence Prediction" task is also employed to jointly pretrain representations for text pairs. It is demonstrated that pre-trained representations minimize the reliance on heavily engineered task-specific architectures. BERT is recognized as the first fine-tuning-based representation model to achieve state-of-the-art performance across a wide range of sentence-level and token-level tasks, surpassing numerous task-specific architectures.

The pre-trained BERT model can be adapted for various tasks, like question answering and language inference, by adding just one output layer, achieving state-of-the-art results without significant changes to its core architecture.

Because of self-attention, BERT can handle different forms as input (single text or pairs of text) by simply changing the input and output layers. This makes the fine-tuning process simple. To finetune BERT to a specific task, you just need to:

- Provide the appropriate input format (e.g., two sentences for paraphrase detection).
- Add a task-specific output layer (e.g., a classification layer for classification tasks).
- Fine-tune all the parameters of the pre-trained BERT model using labeled data for that specific task.

For tasks that require classifying input sequence, the output is derived from the special [CLS] token representation. The [CLS] token is a special token added to the beginning of every input sequence in BERT. Its last hidden state is considered as aggregated sequence representation for classification tasks. This [CLS] representation is fed into an output layer designed for classification.

4.1.1 DNABERT

DNABERT is an early large language model (LLM) for genomes and one of the first applications of transformers to genomics based on the original BERT model.

In genomic research, the context-dependent behavior of cis-regulatory elements (CREs) poses a significant challenge. The same CREs can exhibit diverse functions and activities across different biological contexts (Polysemy of CREs). Additionally, multiple CREs, even those located far apart on the genome, can interact and cooperate, leading to context-specific utilization of alternative promoters with varying functional roles.

DNABERT was designed with the following objectives: (i) to comprehensively consider all contextual information for distinguishing polysemous CREs, (ii) to develop a versatile understanding that can be applied across multiple tasks, and (iii) to achieve strong generalization performance even in scenarios with limited labeled data [12].

In the paper, performance of DNABERT regarding to mentioned objectives is evaluated on some downstream tasks, and the fine-tuned model demonstrates exceptional performance, for instance, in identifying proximal and core promoter regions, significantly surpassing other models in accuracy. It also excels in pinpointing transcription factor binding sites (TFBSs) with high precision, effectively capturing key regulatory sequences crucial for gene expression. This comprehensive proficiency establishes DNABERT as a versatile and powerful tool for advancing genomic research.

DNABERT adopts the same training process as BERT and utilizes the same architecture. This architecture includes 12 Transformer encoder layers, each with 768 hidden units and 12 attention heads, as illustrated in figure 2. Each encoder layer comprises two sub-layers: a multi-head self-attention layer and a fully connected feed-forward layer. To enhance training efficiency, skip/residual connections and layer-wise normalization are integrated around each sub-layer. DNABERT uses multi-head self-attention to better capture relationships between words or elements in a sequence, since it allows the model to process and integrate information from multiple parts of a sequence simultaneously. This enhances the model's ability to capture complex relationships within the sequence, reducing the need for many layers compared to using standard self-attention.

DNABERT processes sequences with a maximum length of 512. For sequences exceeding this limit, the DNABERT-XL approach is suggested by dividing long sequences into smaller segments and concatenating their representations to form the final representation. While DNABERT-XL is a practical solution for extending sequence length support, its efficiency and performance may fall short compared to models explicitly designed for long sequences. Depending on the use case, these limitations may make DNABERT-XL less ideal for processing extremely long genomic sequences.

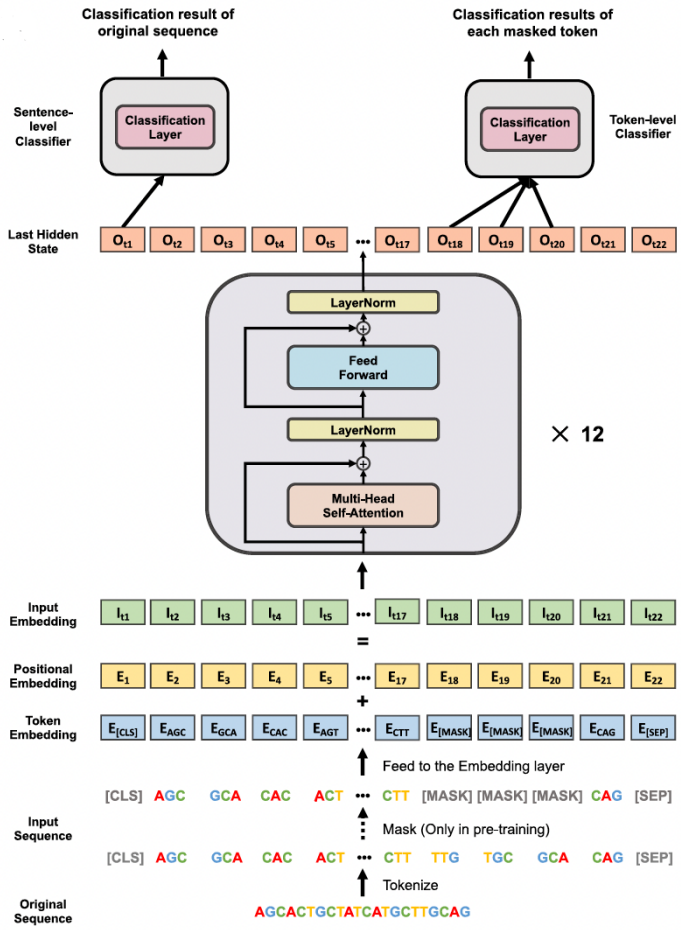


Figure 4.2: DNABERT Architecture, figure from [12].

Tokenization

DNABERT utilized overlapping k-mers for tokenizing DNA sequences. Since different values of k result in varying tokenization of a DNA sequence, by setting k to 3, 4, 5, and 6, four distinct pretrained models based on tokenization are prepared. For each pretrained model, the vocabulary comprises all possible k-mer permutations, along with five special tokens: [CLS] for classification, [PAD] for padding, [UNK] for unknown tokens, [SEP] for separation, and [MASK] for masked tokens. Consequently, the vocabulary size for DNABERT-k is $4^k + 5$.

Pre-training

Pre-training enables the model to learn comprehensive representations of genomic sequence data, capturing both short- and long-range relationships within the sequences. For pre-training DNABERT, two methods were used to generate training data from human genomes. First, the complete genome was split into non-overlapping sub-sequences, and second, random sub-sequences were sampled with lengths ranging from 5 to 510 bp. The input sequence is tokenized

into k-mers, and to mark the beginning and end of the sequence, special tokens were added: [CLS] at the start, representing the entire sequence, and [SEP] at the end. After tokenization, inputs are fed into an embedding layer, including token embedding and positional embedding. Each embedded input is treated as an independent sequence for model training. Only masked language modeling (MLM) tasks were performed during pre-training. Since DNA has a unique grammar in its k-mer representation, a challenge arose in token masking, as masked k-mers could be easily inferred from their neighboring tokens. Overlapping k-mer tokenization ensures that adjacent tokens share $k-1$ characters, leading to significant information leakage in masked language modeling. To mitigate this in DNABERT, contiguous k-length spans of k-mers were masked instead of individual tokens, maintaining the difficulty of the pre-training task. However, partial information leakage remains unavoidable for the leftmost and rightmost masked tokens. The model was trained using the last hidden state of each masked token, and a cross-entropy loss was applied over all masked k-mers.

Fine-tuning

Most DNA-related applications can be conveniently categorized into two types: sequence-level tasks and token-level tasks. For instance, DNA replication origin prediction can be considered as a sequence-level 2-class classification task, where class 1 represents "the given DNA sequence contains a replication origin," and class 2 indicates the absence of one. Additionally, masked token prediction can be treated as a token-level V-class classification task, where V corresponds to the vocabulary size, and each class represents a unique token within the vocabulary. By fine-tuning the model by task-specific data, DNABERT can address both the token-level and sequence-level tasks.

4.1.2 DNABERT-2

DNABERT-2 builds on the Transformer Encoder architecture, like its predecessor DNABERT. However, it introduces significant improvements by pre-training a foundational model designed to accommodate multi-species genomes, including human, mouse, yeast, and virus sequences. This multi-species adaptability is intended to improve DNABERT-2's versatility and allow it to handle diverse genomic contexts and broaden its application to various species and sequence prediction tasks.

DNABERT's overlapping k-mer tokenization strategy presented challenges during pre-training, including information leakage and reduced computational efficiency due to the relatively long tokenized input sequences. Specifically, in masked language modeling, the model ideally selects the best option from the entire vocabulary, thereby learning to differentiate between numerous possibilities. However, information leakage, caused by the overlapping k-mers, reduces this search space. This artificially simplifies the task, leading to poor sample efficiency, as the model is not adequately challenged to learn truly meaningful patterns. Furthermore, a length L input sequence yields L - k + 1 overlapping k-mer tokens, each of length k, introducing significant redundancy. This redundancy, coupled with the quadratic computational complexity inherent in Transformer-based models, results in low training efficiency. To address these challenges, DNABERT-2 applied an alternative tokenization approach to over-

come the limitations of k-mer tokenization. Specifically, it replaced overlapping k-mer tokenization with Byte Pair Encoding (BPE) [35], a statistical data compression algorithm. BPE constructs tokens of variable-length by iteratively merging the most frequently co-occurring genome segments in the dataset. The aim of this strategy is eliminating information leakage but also benefits from the computational efficiency of non-overlapping tokenization, enhancing scalability and overall model performance. A vocabulary size of 4096 was selected, including tokens with various sizes. This vocabulary is used in the DNABERT-2 model for tokenization, as it was determined to provide the optimal balance between model performance and computational efficiency among the evaluated candidates.

Furthermore, DNABERT-2 introduced enhancements to improve the model's efficiency and capabilities, including: Replacing learned positional embeddings with Attention with Linear Biases (ALiBi) [36] to address the input length constraint; Leveraging Flash Attention [37] and Low Precision Layer Normalization to optimize computational and memory efficiency. And incorporating Low-Rank Adaptation (LoRA) [38] during the fine-tuning stage (if needed) to enable parameter-efficient training.

4.2 Fine-Tunning DNABERT Models for ORIs Prediction

In genomics research, as in many other fields of study, unlabeled data is typically far more abundant than labelled data. In this study, we tried to leverage the concept of transfer learning using genome language models to capture general patterns and representations of the genome during the pre-training phase, then fine-tune for identifying DNA replication origins using labelled data. Thus, it takes advantage of the broad generalization skills developed during pre-training and enables the model to effectively adapt to the specific task with less labelled data by fine tuning. Hence, for our downstream task, we used pr-trained DNABERT models and fine-tuning to discriminate origin from non-origin sequences directly from the final embedding of the sequence alone. This approach eliminates the need for labor-intensive feature extraction processes commonly required in traditional machine learning methods discussed in Chapter 2.

Since our primary objective in this research is to unravel the DNA base composition underlying replication origins, rather than solely using models as AI tools for prediction, our approach includes data engineering and explainability as well.

4.3 Data Engineering

As mentioned in chapter 3, replication origins constitute a small portion of the entire genome, making the subsampling of non-origin sequences challenging because it directly impacts model ability to differentiate origins from non-origins. Our work adopts a data engineering approach to tackle this challenge by creating multiple datasets that share the same positive instances but include different negative instances varying in difficulty and properties. This enables us to systematically compare the performance of models fine-tuned on these datasets,

providing deeper insights into the discriminative properties that characterize replication origin sequences. Additionally, it sheds light on how applied language models handle varying levels of complexity, offering a more comprehensive understanding of their capabilities for this specific task. In chapter 3, the procedure of subsampling each dataset and the main objectives of these datasets are described in detail. We used DNABERT and DNABERT-2 pretrained models to fine-tune all datasets, and results are discussed in Chapter 5.

4.4 Explainability

There are two primary approaches to explain large language models in the explainability literature: local explanations and global explanations. Local explanations focus on clarifying how the model makes a prediction for a specific input instance, while global explanations aim to provide a comprehensive understanding of the model's overall behavior and functioning [39].

In this study, we focus exclusively on local explanations, as the objective is to determine whether pretrained language models can identify the known properties of replication origin sequences, and if their predictions may be influenced by other factors as well. Local explanations encompass various approaches for generating explanations; however, we focus on two specific methods: attention-based explanation and perturbation-based explanation, which is a kind of feature attribution-based explanation according [39].

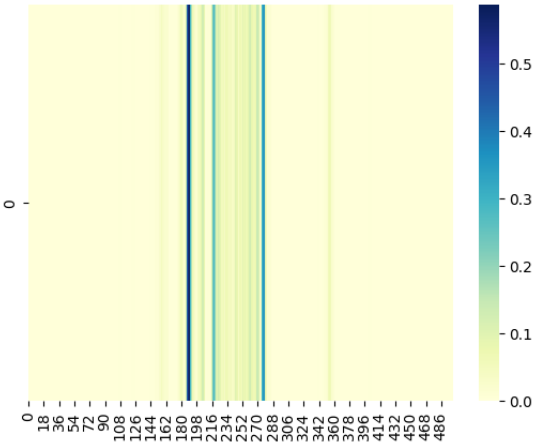
4.4.1 Attention-based Explanation

Transformer attention scores have been suggested as an interpretability approach to address the challenge of black-box models in genomics. As a result, some transformer-based models have directly presented analyses of their attention scores as evidence of their capacity to capture biological signals. The attention mechanism is often regarded as a tool to focus on the most relevant input components, potentially capturing meaningful correlations that explain model predictions. Many explanation methods leverage attention weights or analyze the knowledge encoded in attention, which can be broadly grouped into visualization, function-based, and probing-based techniques. Here visualization is selected as attention-based explanation. Despite the extensive debate in research regarding the suitability of attention weights for explanations, we found it valuable to assess whether they can effectively explain the performance and behavior of DNABERT models.

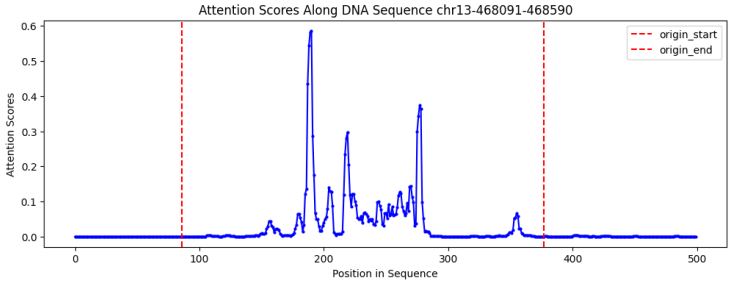
By means of the self-attention mechanism, the DNABERT model demonstrated significant accuracy in classifying origin and non-origin sequences. Building on this, we applied an attention-based explanation to identify key motifs that the model is considering as evidence for its final classification decisions. To achieve this objective, we applied two approaches, which are described in the following sections.

In DNABERT, a module is provided for extracting attention scores from the last hidden state and visualizing scores in nucleotide-level directly. In this module, the fine-tuned model is first loaded. Each input sequence is then fed into the model to extract attention scores from the final layer. These scores are represented by eight matrices, each corresponding to an attention head.

Each matrix has dimensions of N × N, where N is the length of the tokenized input sequence. These scores capture high-level features learned from the input sequence. In BERT-style models, the last hidden state of the [CLS] token serves as a global representation of the entire input sequence. By extracting the rows corresponding to the [CLS] token from the attention score matrices, we obtain its attention distribution across all tokens. Summing over all attention heads results in a single attention score vector per token. Since DNABERT tokenizes sequences using overlapping k-mers, we compute attention scores at the base pair level rather than the k-mer level. To achieve this, the attention scores of all k-mers containing a specific base pair are averaged, providing a refined per-base attention score. These attention scores highlight the regions of the sequence that the [CLS] token attends to the most and can be visualized along the sequence using a heatmap or a line plot as illustrated in Figure 4.3.



(a)



(b)

Figure 4.3: Example of attention scores visualization corresponding to the last hidden state of [CLS] token in DNABERT, a) heatmap, b) line plot representation for same sequence.

In Figure 4.3, we observe that attention scores within the actual replication origin range are significantly higher than those outside this region, exhibiting multiple distinct peaks. In contrast, the scores outside this range remain relatively flat and close to zero. Interestingly, this pattern is consistently observed

across most origin samples, allowing us to focus on high-attention regions to enhance model explainability and assess its ability to recognize biologically relevant motifs.

DNABERT's Explainability tool

As an initial step, we utilized the motif analysis tool provided by DNABERT to identify significant motifs based on attention score distributions. The module steps are as follow:

- Contiguous high-attention regions within input sequences were identified based on 3 conditions, which are defined with suggested cutoff values as: i) attention be greater than mean of attention within the sequence; ii) attention must be greater than X*(minimum attention) when X=10; and iii) contiguous regions must have a minimum length of L=4.
- Selected regions were treated as preliminary motif instances, and enrichment in positive sequences was tested using a hypergeometric distribution. Statistical significance was assessed using a hypergeometric test with Benjamini-Hochberg correction (P-value$<$ 0.005), resulting in the identification and retention of key motifs while filtering out less significant ones.
- Selected motifs are analyzed using pairwise alignment, and similar motifs are merged into the same group. The output is a dictionary where the keys represent the general motif forms, and the values correspond to the sequence index and the motif's position within the sequence.
- Since final motifs may have different lengths, motifs were converted to a uniform window length (user defined) around the center of each motif instance.
- Final motifs were converted to Weblogo format.

For our task, using default cutoff values did not yield robust motifs. Since we need to find longer motifs, we tailored the analysis to our specific problem by adjusting the cutoff values (X=5, L=8, P-value$<$ 0.05). As a result, we identified several significant motifs that will be discussed in detail in the next chapter.

Although the motifs identified by this tool were closely aligned with ACS motifs or A/T-rich motifs in the flanking regions of ACS, they lacked robustness in terms of p-value, and length. This limitation likely stems from its reliance on continuous high-attention regions as the initial search space. Since the model does not always assign high attention across the entire continuous region, significant portions of motifs may be missed by the algorithm. For instance, in Figure 4.3 b, which illustrates attention scores of an origin sequence, multiple sharp peaks with high attention exist, while the surrounding peak typically falls below average attention score. Consequently, selected motifs tend to be too short, even when using more relaxed cutoffs. Furthermore, the tool's lack of advanced clustering strategies, such as hierarchical or multiple sequence alignment, resulted in fragmented and less comprehensive motif representations. To address these limitations, we proposed an alternative motif discovery pipeline based on attention score visualization, which is detailed as follows.

Proposed Pipeline for Motif discovery

Building upon the promising results obtained from the motif analysis tool in DNABERT, we propose an alternative approach for motif discovery. We extracted attention scores for the [CLS] token, from the last hidden layer of our fine-tuned model(Figure 4.4, a). By visualizing attention plots for origin and non-origin sequences, we identified a recurring pattern. Within the origin range, multiple sharp peaks with high attention scores were observed, as illustrated in Figure 4.4, b. To enhance motif robustness, we defined the search space by focusing on subsequences surrounding these peaks.

To enhance the robustness of attention-based motif discovery, we refined the search space by concentrating on 20 bp fragments around high-attention peaks. For each input sequence, we selected up to four top peaks based on an attention score threshold, ensuring that the total length of selected fragments stayed below 16% of the sequence. Only peaks separated by at least 10 bp were included, limiting the fragment overlap to a maximum of 10 bp. High-attention 20 bp fragments were extracted from both train and test sets and independently

Subsequently, to perform a comprehensive search and motif discovery, we leveraged a more advanced bioinformatics tool, MEME (Multiple EM for Motif Elicitation) [40]. The MEME algorithm is a motif discovery algorithm using the Expectation-Maximization (EM) algorithm to fit a mixture model to sequences of biological sequences (such as DNA and proteins). The method models motifs as unknown probabilistic patterns and iteratively refines their representations using a mixture model, where the EM algorithm optimizes the likelihood of observed sequences. This unsupervised learning technique does not require prior knowledge of motif locations, making it highly adaptable for discovering regulatory elements in DNA, RNA, and protein sequences.

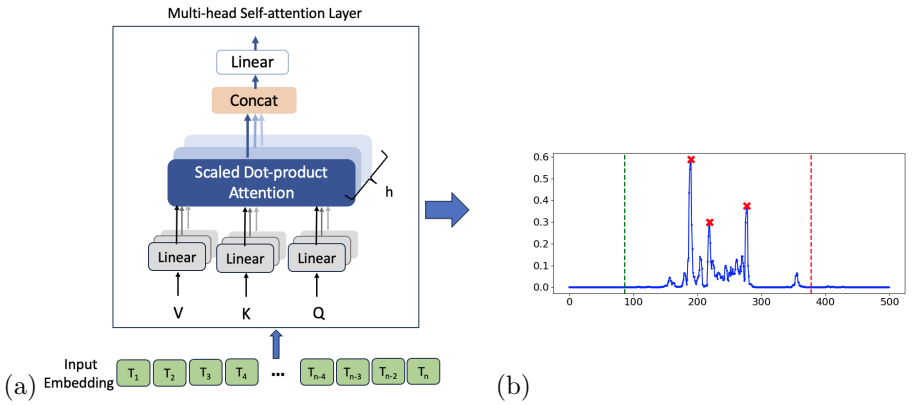


Figure 4.4: (a) Attention scores being extracted from the multi-head attention layer, (b) Attention scores across an origin instance. Motif discovery targets are fragments of 20 bps around peaks (red crosses) of the attention scores (values along the Y axis). Vertical lines correspond to origin start (green) and end (red) with sequence position along the X axis.

The goal of this approach is to leverage high-attention regions for motif discovery effectively, enabling an assessment to what extent DNABERT's dis-

criminative power is driven by its ability to identify biologically meaningful motifs.

This approach was evaluated on the Rand-Neg and ACS-Neg datasets, using models fine-tuned for each, since they better represent real-world origin prediction with diverse non-origin samples. In contrast, non-origin samples in Shuffled-Neg and Block-5-Shuffled-Neg were artificially generated through shuffling. While performance analysis of DNABERT on these datasets is important for understanding model learning, the results of the motif analysis may be misleading, and therefore we omit them.

For the model fine-tuned on the Rand-Neg dataset, we applied extracted high-attention peak's subsequences for both the test set and train set, to process independently. First, we applied MEME motif discovery in classic mode exclusively to positive samples. Then, MEME in discriminative mode is used to identify motifs that distinguish between the two classes by considering fragments from both positive and negative samples.

We also applied the same procedure to the ACS-Neg dataset, where both negative and positive instances contain at least one ACS motif. Since the model fine-tuned on this dataset has a moderate performance, the objective was to identify motifs beyond ACS motifs that could serve as discriminative features. A detailed discussion on the explainability of the results are presented in Chapter 5.

Attention Explainability in DNABERT-2

For DNABERT-2, attention score extraction was not readily available, so we developed a custom module for this purpose. This tool extracts attention scores from the last hidden layer of the DNABERT-2 model, which has been fine-tuned on our origin datasets. Extracted attention scores are in token levels that need to be presented also per nucleotide. Since DNABERT-2 uses Byte Pair Encoding (BPE) tokenization, which produces non-overlapping tokens of varying lengths, the attention score assigned to each token by the model is uniformly distributed across all base pairs within that token.

At first, we tried to visualize attention scores of the last hidden state of the [CLS] token (average over all attention heads). In BERT-style models, the [CLS] token aggregates global sequence information, highlighting the regions of the sequence that the [CLS] token attends to the most. But this representation is for DNABERT-2 biased toward the beginning of the sequence (to itself and its neighboring tokens) and diminishing toward the sequence end, as illustrated in Figure 4.5. Therefore, the attention scores from [CLS] cannot directly reflect important regions, making it unsuitable for motif discovery.

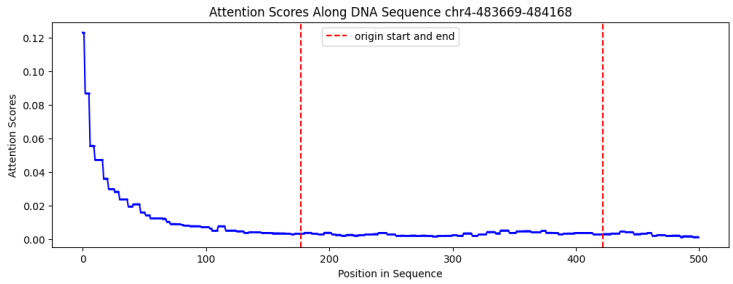


Figure 4.5: Attention scores corresponding to the [CLS] token extracted from DNABERT-2

To enhance the interpretability of the attention map, we explored an alternative representation by visualizing token-level attention. Instead of relying solely on the [CLS] token, we extracted attention scores from the final layer for all tokens and averaged them. This approach resulted in a different visualization, as shown in Figure 4.6, for the same sequence depicted in Figure 4.5. However, this representation also does not clearly highlight high-attention regions, making it more challenging to pinpoint potentially important sequences for motif discovery.

Moreover, unlike DNABERT, DNABERT-2 does not exhibit distinctly higher attention scores within the replication origin range compared to regions outside it (as seen by comparing Figure 4.3 with Figure 4.5 and 4.6)). This suggests that DNABERT-2 may not prioritize origin-related motifs as strongly, which could impact its ability to leverage attention scores for identifying biologically relevant motifs.

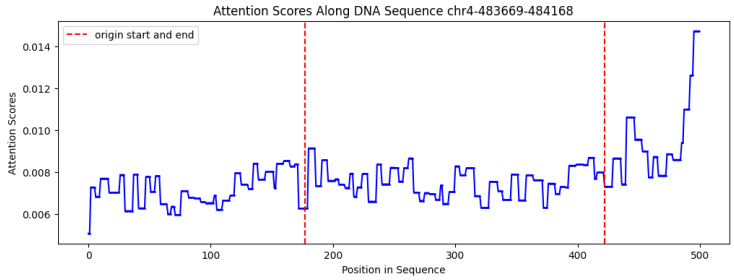


Figure 4.6: Average attention scores over all tokens extracted from DNABERT-2

For this model, observing unclear high-attention regions and the difficulty in identifying important motifs, aligns with the ongoing debate in the explainability literature regarding attention mechanisms. Some studies argue that attention mechanisms do not necessarily capture the structure or relationships within input sequences, as in NLP, this is observed in the failure to capture syntactic structures [39]. It is also argued that the raw attention often contains redundant information, reducing its reliability to reflect true feature importance. Therefore, mentioned attention representations for DNABERT-2, align with the criticism that observing attention scores may not reflect true feature importance

visually.

However, different visualization systems vary in how they depict relationships across multiple scales, so we can use various representations of attention for different models. A widely used method for visualizing attention scores is two dimensional heatmaps, which display attention as a matrix. In Figure 4.7, an example of this representation is shown. Here, we extracted attention scores from the last layer of the fine-tuned model for a single sequence and summed scores across all attention heads, resulting in a matrix of attention scores. This matrix represents the attention scores between tokens, where each element (i, j) indicates the degree to which token i attends to token j. Higher values signify stronger attention, meaning the model places greater importance on those token relationships. In chapter 5, we analyzed heatmaps of DNABERT-2 and DNABERT, to interpret how models process sequences, and compare attention mechanisms across two models.

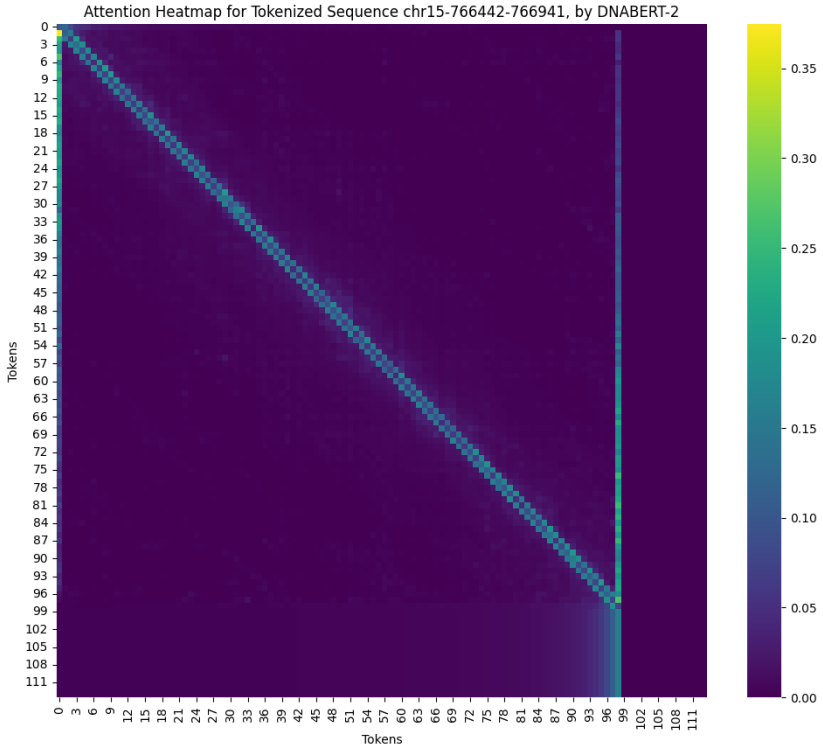


Figure 4.7: Attention heatmap representation for DNABERT-2

4.4.2 Perturbation-Based Explanation

Perturbation-based explanation methods evaluate a model's decision-making process by altering or removing parts of the input and observing the effect on the output. The simplest form, leave-one-out, removes or modifies specific input features — for example by masking specific tokens — and measures their importance based on the resulting prediction changes. These methods aim to identify the smallest subset of features that significantly alters the model's prediction,

with feature importance quantified using metrics like confidence scores or advanced techniques like reinforcement learning [39]. While perturbation-based methods provide a straightforward and interpretable way to assess feature importance, their effectiveness depends on addressing challenges like feature correlations, overconfidence. However, we applied this method to evaluate the importance of ACS motifs in predictions provided by the fine-tuned DNABERT-2 model. To achieve this objective, we evaluated true positive samples containing at least one ACS motif. By systematically removing these motifs and reintroducing the modified samples into the fine-tuned model, we analyzed the extent to which the classification probability was affected. This enables us to explain the extent to which the model's prediction relies on ACS motifs.

For DNABERT-2, which uses BPE tokenization to produce non-overlapping tokens, an alternative strategy for perturbing samples involves randomly shuffling the tokenized sequences. By feeding shuffled samples into the fine-tuned model, we analyzed the impact of this perturbation on the classification probability to assess the model's sensitivity to token order.

Chapter 5

Results

This chapter presents the results of our study, focusing on both the performance and explainability of the models evaluated. We applied pretrained DNABERT and DNABERT-2 and fine-tuned them independently on our datasets to compare the performance of each model across multiple datasets with various levels of complexity.

First, performance analysis provides a detailed assessment of the predictive capabilities of evaluated models for each dataset. Then, we aim to interpret the presented results and evaluate the model's performance in our downstream task using the explainability approaches introduced in Chapter 4. Explaining the results offers valuable insights into the model's effectiveness in leveraging attention mechanisms to capture biologically relevant patterns.

5.1 Performance Analysis

We evaluated both models on the four datasets described in Section 3.2. To ensure robust and reliable results, we utilized seven different data splits for each dataset, fine-tuning the model separately on each split, resulting in seven fine-tuned models per dataset. The performance was assessed on the corresponding test sets of these seven splits, and the average performance across all splits is reported for each dataset. Based on objectives of each dataset, we have discussed the model behavior when it is finetuned on multiple datasets including different negative samples (non-origin sequences), vary in difficulty and properties.

It is worth mentioning that our four datasets can be categorized into two groups: those with negative subsampling using actual genome fragments, including the Random-Neg and ACS-Neg datasets, and those with artificially generated negative samples, such as the Shuffled-Neg and Block-5-Shuffled-Neg datasets. When comparing model performance, it is more reasonable to evaluate datasets within the same category, considering their main objective of differing strategies for subsampling.

5.1.1 DNABERT Performance

For DNABERT, pretrained models were available for four different tokenization schemes. We finetuned models on 3-mer, 4-mer, and 6-mer tokenization and

found their performance to be similar for our task. Therefore, we only report the results of 4-mer tokenization for each dataset in Table 5.1. Results presented in this table are average performance over seven data splits, and dataset splitting strategy was random splitting.

The Random-Neg dataset, with an average accuracy of 0.83 and area under the curve (AUC) of 0.90 on test set, demonstrates significant discriminative power when non-origin sequences are randomly selected from the genome. However, ACS-Neg dataset, performance decreased compared to the Random-Neg dataset. This outcome was expected, as origin and non-origin sequences in ACS-Neg dataset share greater similarity due to the presence of ACS matches in both, which make discrimination more challenging. However, the model was still able to distinguish between them with an accuracy of 0.74, suggesting that it leverages other discriminative features beyond ACS matches. To explore these features, we conducted a motif discovery analysis, which is discussed in Section 5.2.1.

To compare the ability of model in capturing local (short-range) and global (long-range) dependency in DNA sequences, we can compare the model's performance when fine-tuned on the Shuffled-Neg and Block-5-Shuffled-Neg datasets.

DNABERT achieved its best performance when fine-tuned on the Shuffled-Neg dataset, yielding an accuracy of 0.90 and an AUC of 0.96. This dataset comprises non-origin samples generated by shuffling the nucleotides of origin sequences, thus preserving the overall nucleotide frequencies (A, T, C, G) in both classes while disrupting their order. These results suggest that ordered nucleotide patterns are crucial for distinguishing origin sequences. The disruption of both local and global sequence patterns in the non-origin sequences creates a more distinct boundary between the two classes, enhancing the model's discriminative power. This also underscores the ability of BERT-style models to capture both local and global sequence dependencies, specificity on our task including the ordered characteristic patterns of replication origin sequences.

On the Block-5-Shuffled-Neg dataset, where local sequence patterns are preserved but global patterns are disrupted, the model faced a more challenging classification task. Compared to the Shuffled-Neg dataset, performance decreased to 0.77 accuracy and 0.86 AUC. This drop in performance suggests that long-range sequence patterns also play a significant role in DNABERT's ability to distinguish origin sequences.

Table 5.1: DNABERT Performance on all datasets

Dataset	Accuracy	AUC	Precision	Recall
Random-Neg	0.83	0.90	0.83	0.83
ACS-Neg	0.74	0.82	0.74	0.74
Shuffled-Neg	0.90	0.96	0.90	0.90
Block-5-Shuffled-Neg	0.77	0.86	0.77	0.77

To assess the reliability of random data splitting and ensure it does not lead to overestimation of performance, we applied chromosome-based data splitting to the Random-Neg dataset. The model's performance on this dataset was slightly better than that obtained with random splitting, indicating that the random splitting did not bias the model's performance. The average performance across seven different chromosome-based splits is presented in Table 5.2.

Table 5.2: DNABERT performance, chromosome-based data splitting

Dataset	Accuracy	AUC	Precision	Recall
Random-Neg	0.85	0.91	0.85	0.85

5.1.2 DNABERT-2 Performance

For DNABERT-2, we fine-tuned the model on all datasets. The results shown in Table 5.3, represent the average performance on test set across seven randomly dataset splits.

Table 5.3: DNABERT-2 Performance on all datasets

Dataset	Accuracy	AUC	Precision	Recall
Random-Neg	0.81	0.82	0.83	0.81
ACS-Neg	0.75	0.72	0.76	0.75
Shuffled-Neg	0.92	0.92	0.93	0.92
Block-5-Shuffled-Neg	0.83	0.83	0.84	0.82

DNABERT-2 exhibits lower performance compared to its predecessor on the Random-Neg dataset. Specifically, DNABERT achieves 0.83 accuracy and 0.90 AUC, while DNABERT-2 drops to 0.81 accuracy and 0.82 AUC. On the ACS-Neg dataset, performance decreased as expected compared to the Random-Neg dataset. However, the results still demonstrate considerable discriminative power, even when both classes contain ACS motifs. To compare two model on ACS-Neg dataset, DNABERT-2 demonstrates comparable accuracy but lower AUC than DNABERT. These results suggest that DNABERT possesses greater discriminative power for distinguishing replication origin sequences from non-origin sequences derived from actual genome data.

For the Shuffled-Neg dataset, DNABERT-2 performs comparable as preceded model, achieving an accuracy of 0.92 and an AUC of 0.92, compared to accuracy of 0.90 and AUC of 0.96. For the Block-5-Shuffled-Neg dataset, DNABERT-2 shows better accuracy than its predecessor, 0.83 vs. 0.77. These results suggest that both models are effective in capturing short-range dependency, but DNABERT-2 may be less effective in capturing long-range dependency. This is evident from its less drop in accuracy compared to DNABERT when it is finetuned by Block-5-Shuffled-Neg datasets instead of Shuffled-Neg dataset.

We applied chromosome-based data splitting to Random-Neg on DNABERT-2 as well. The results showed that performance of DNABERT-2 was comparable to that of the Random-Neg dataset with random splitting, indicating that random splitting does not significantly bias the model's performance.

Table 5.4: DNABERT-2 performance, chromosome-based data splitting

Dataset	Accuracy	AUC	Precision	Recall
Random-Neg	0.79	0.80	0.81	0.78

5.2 Explainability of Results

5.2.1 DNABERT

We applied DNABERT's explainability tool to extract motifs based on attention scores. The model fine-tuned on the Random-Neg dataset is used here, since it demonstrated robust performance and closely aligned with real-world origin prediction challenges, by ensuring diversity in non-origin samples. As a result, several significant motifs were identified from high-attention regions under the conditions described in Section 4.6.1.1. The most significant motifs are shown in Figure 5.1. Some of these motifs, such as (a) and (b), are considerably aligning with the ACS motif scheme 5'-WTTTATRTTTW-3', presented in Table 1.1. While some other motifs can be aligned by considering their reverse complement sequences, such as (d) and (e). Also, certain motifs consisting of continuous A or T bases, are found, like (f) and (h). As discussed in Section 1.3, these repetitive patterns are commonly found in intergenic ORIs and upstream and downstream of replicative ACSs.

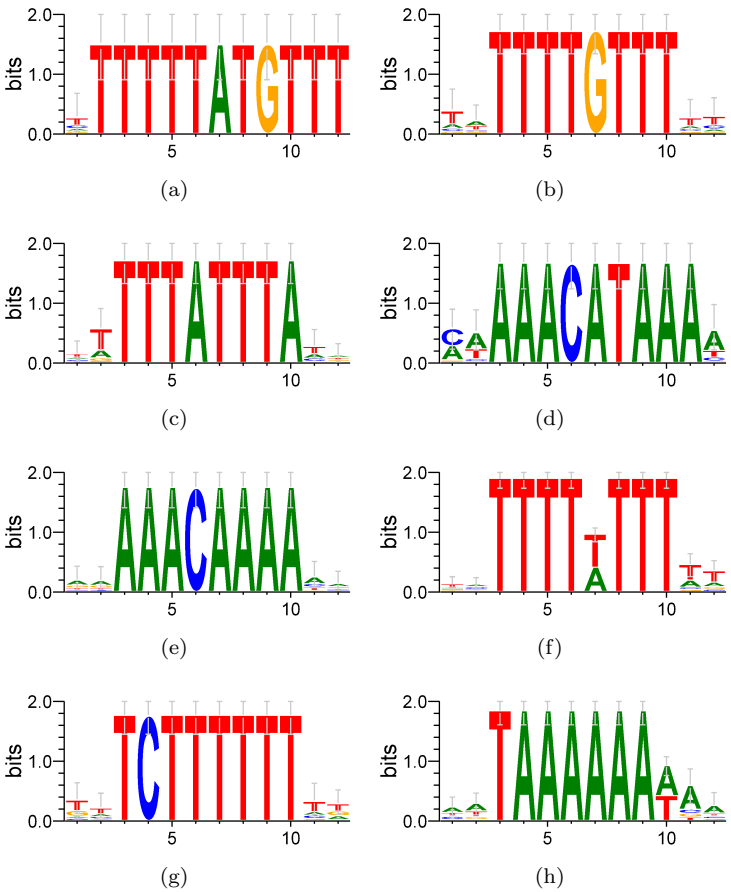


Figure 5.1: Motifs found by explainability tool of DNABERT

Results of Proposed Motif Discovery Pipeline:

Random-Neg dataset: For the model fine-tuned on the Rand-Neg dataset, we applied our proposed motif discovery pipeline, described in Section 4.4.1 . First, we consider true positives (TP) of the test set, which consists of 65 origin and 65 non-origin samples, where TP = 51. High-attention fragments (20 bps) from TP cases were analyzed using MEME in the classic mode. As shown in Figure 5.2, the motif discovered using this approach is highly aligned with the ACS motif pattern 5'-WTTTATRTTTW-3', with an E-value of 2.6e-006, indicating strong statistical significance.

Figure 5.2: Motif found for true positive samples: TTTTTWTTTATRTTT ,E-value: 2.6e-006

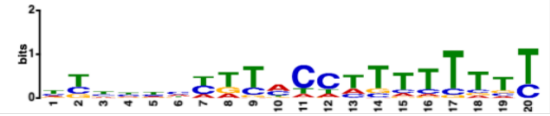
We also applied this method to the training set of the Random-Neg dataset, which comprises 228 origin and 227 non-origin samples. Initially, motif discovery was conducted using MEME in classic mode on a set of subsequences derived from high-attention peaks in all positive sequences (origin samples). Results are shown in Figure 5.3. Subsequently, MEME was applied in discriminative mode, incorporating two sets; subsequences from high-attention peaks in both positive sequences and negative sequences, independently. This approach aimed to identify motifs that are not only frequent in origin sequences but also distinct from non-origin sequences, to do more comprehensive motif discovery. These motifs are depicted in Figure 5.4. The identified motifs illustrated consistently exhibited strong similarity to the ACS motif pattern, also continuous A or T bases, commonly found in intergenic ORIs and upstream and downstream of replicative ACSs. These results can be consider as further validating the effectiveness of DNABERT in capturing biologically relevant patterns. The difference between motif found for test set and motifs for train set, stems from greater variability in the training set, resulting in broader motif representations than in the test set.

(a)

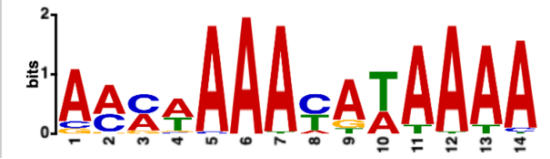
(b)

Figure 5.3: Motif found for train set origins, discoverd by MEME classic mode. a) 'TATATTTATRTWTWT', E-value: 2.3e-032, b) Motif 'RAAAAAAAG-GKAARGNRAAV', E-value: 2.5e-014

(a)



(b)



(c)

Figure 5.4: Motif found for train set, discoverd by MEME discriminative mode. a) Motif 'ATATWTATATWTATRTWTWT', E-value: 1.4e-037, b) 'TYYYYMYTTMCCTTTTTTTT', E-value: 4.0e-031, c) 'AMMWAAAYAWAAAA', E-value: 3.7e-014.

ACS-Neg dataset: To identify more discriminative motifs beyond ACS motifs, we applied our pipeline to the ACS-Neg dataset. Both classic and discriminative modes were tested on the training dataset. The results, presented in Figure 5.5, reveal two statistically significant motifs in E-value. These motifs exhibit patterns characterized by alternating A or T bases, which are commonly associated with intergenic ORIs, as discussed in Section 1.3.

(a)

(b)

Figure 5.5: Motifs found for train set, a) by MEME classic mode 'ATATATATRTR', E-value = 5.7e-059, b) by MEME discriminative mode 'TWTATATRTATDTRT', E-value: 2.4e-074.

5.2.2 DNABERT-2

Attenion-Based Explainability:

As mentioned in section 4.4.1, for DNABERT-2 attention maps do not reveal clearly defined and distinct high-attention regions, so it is challenging to identify potentially important ranges for motif discovery. Therefore our proposed motif discovery pipeline which is depended on extract motifs from subsequences around high peaks was not effective here. However, we tried to find motifs by adapting DNABERT find motif tool for DNABERT-2. By adjusting loser cutoff values for selecting high-attention regions, we extracted some motifs. While this tool identified A/T-rich motifs, they were relatively short and less robust compared to those identified by DNABERT attention scores. Some of the most frequent are illustrated in Appendix A.

Analyzing Attention Heatmaps: In Section 5.1.2, by comparing the performance of both models on the Shuffled-Neg dataset and the Block-5-Shuffled-Neg dataset, we suggested that DNABERT-2 seems to be less effective than DNABERT in capturing long-range dependencies. Analyzing the attention heatmaps provides further evidence to support this claim.

As illustrated in Figure 5.6, the distinct attention patterns between two models for same sequence, suggest a fundamental difference in how each model processes token relationships, despite being fine-tuned on the same dataset (Rand-Neg). DNABERT exhibits more non-diagonal attention, meaning that tokens attend not only to their immediate neighbors but also to distant tokens. This indicates that DNABERT effectively captures long-range dependencies, enabling interactions between distant regions in the sequence.

In contrast, DNABERT-2 exhibits stronger diagonal attention, where tokens primarily attend to themselves and nearby tokens. This indicates a greater focus on localized context modeling, with reduced emphasis on long-range interactions. Consequently, this suggests that the model relies more on optimizing

token weighting for classification rather than capturing broader sequence dependencies.

To provide a more detailed explanation for special tokens in the heatmap of DNABERT-2; here the first column represents the [CLS] token, and column 98 is representng [SEP]. Additionally, dark columns appearing after [SEP] correspond to [PAD] tokens added to end of sequence to provide all tokenizad inputs with same length. Notably, partially highlighted columns [CLS] and [SEP] is related to distanc of tokens from beginning and end of sequence, the closer a token is to the special tokens [CLS] and [SEP], the more attentions them.

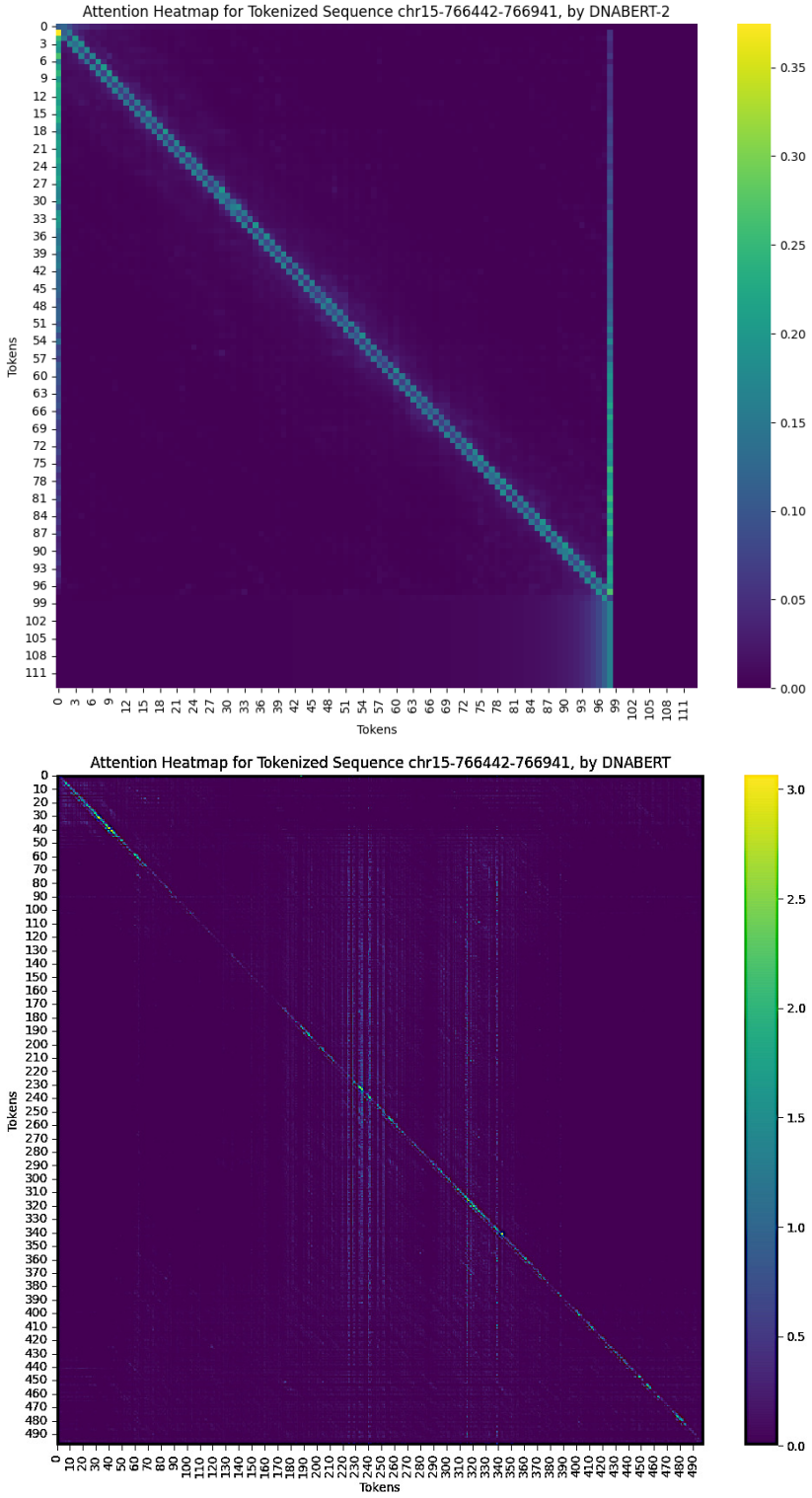
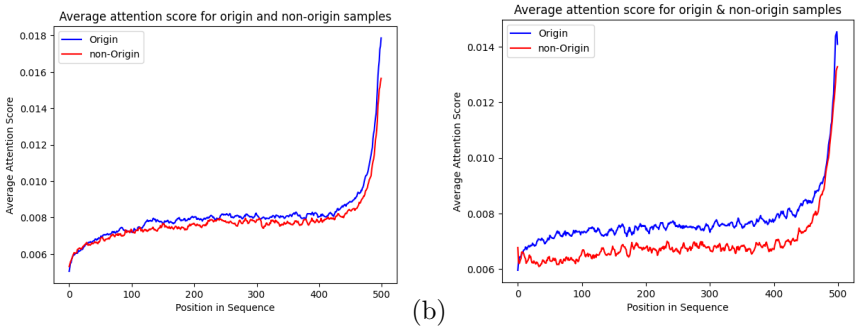


Figure 5.6: Attention maps comparison for same sequences, extracted from fine-tunned DNABERT-2 and DNABERT.

In another attempt to analyze attention scores, we visualize the attention distribution by averaging attention scores across all tokens, as shown in Figure 4.6. By comparing the average attention scores across all sequences between models fine-tuned on the Rand-Neg and Shuffled-Neg datasets, we observe that the gap between average scores of origin and non-origin samples is more in the Shuffled-Neg dataset than in the Rand-Neg dataset.

Since the Shuffled-Neg model outperforms the Rand-Neg model (accuracy: 0.92 vs. 0.81), this suggests that as the model improves in distinguishing between the two classes, the distance of average attention scores between two classes increase. This further supports the idea that the model relies more on optimizing token weighting for classification.



(a) (b)

Figure 5.7: Average attention score comparison of origin and non-origin samples, in DNABERT-2. a) model finetunned on Rand-Neg dataset, b) model finetunned on shuffled-Neg dataset

Perturbation-based Explainability:

Here, we utilize token-level attention visualization as it provides a more informative representation for our purpose. As described in Section 4.4.1, rather than relying solely on the [CLS] token, we extract attention scores from the final layer for all tokens and compute their average. This approach effectively aggregates information by averaging across the columns of the attention matrix (as shown in Figure 5.6), thereby reducing its dimensionality from two to one. First, we visualized attention maps in line plot representation, as illustrated in Figure 5.8, and highlighted intersects with ACS motifs to observe how the model attends to these motifs. The attention score pattern does not show a significant difference compared to the surrounding regions.

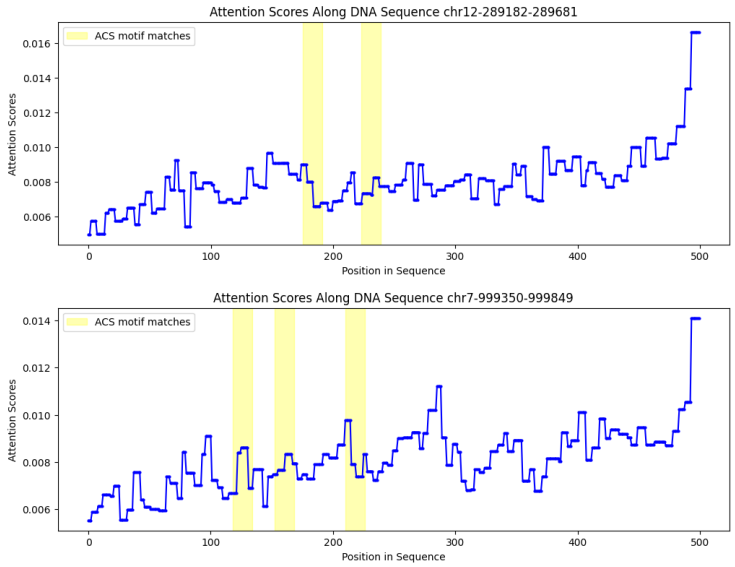


Figure 5.8: Attention score visualization for samples with ACS motif matches

Therefore, we employed a perturbation-based explanation to assess the importance of ACS motifs as a prominent feature in model predictions. As described in Section 4.4.2, we modified true positive samples from the test set by removing the ACS motif while extending the sequence from both the left and right to maintain the original length. We observed that the model's predictions changed for 10% of the samples, leading to their reclassification as non-origin sequences. This suggests that ACS motifs play a role in the model's decision-making process, despite not exhibiting visually distinguishable high scores in the attention score plots. We also visualized attention maps of original sequences alongside their corresponding perturbed sequences, analyzing both cases where classification changes after perturbation and where it remains unchanged. This comparison provides insights into how attention patterns are affected by perturbation. In samples where the model alters its classification post-perturbation, we observe significant changes in attention maps compared to the original sequences. For instance, Figure 5.9 illustrates two origin sequences that, after perturbation, were misclassified as non-origin sequences.

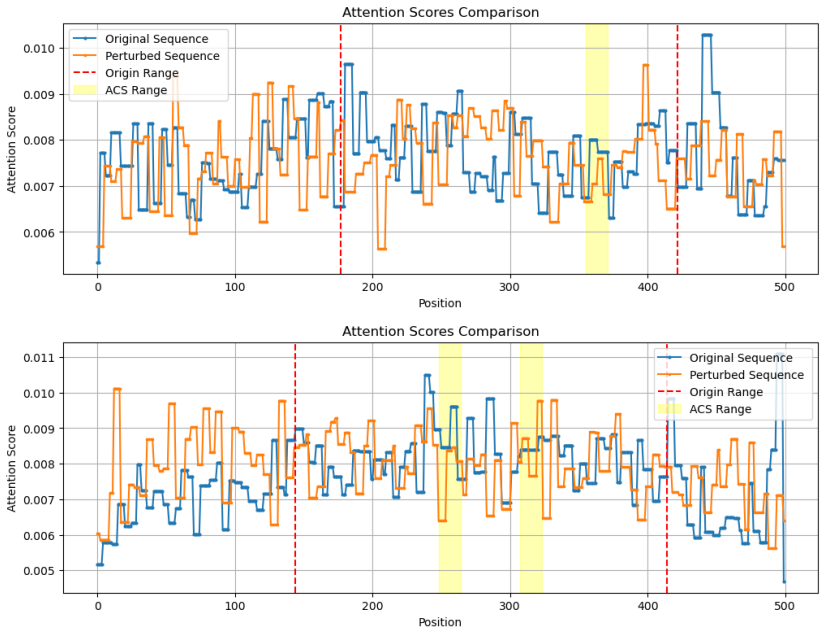


Figure 5.9: Attention score comparison for two samples after ACS motif elimination, leading to a change in model prediction.

Conversely, in cases where perturbation does not affect the model's predictions, the attention maps exhibit a consistent pattern, Figure 5.10. Specifically, for tokens that shift due to the removal of ACS, the attention map merely shifts toward the eliminated motif, indicating that the same tokens retain their original attention scores. Meanwhile, for tokens that remain fixed in their position after ACS removal (between two ACS of Figure 5.10), the attention map remains nearly identical to the original, further reinforcing the stability of attention patterns in these instances.

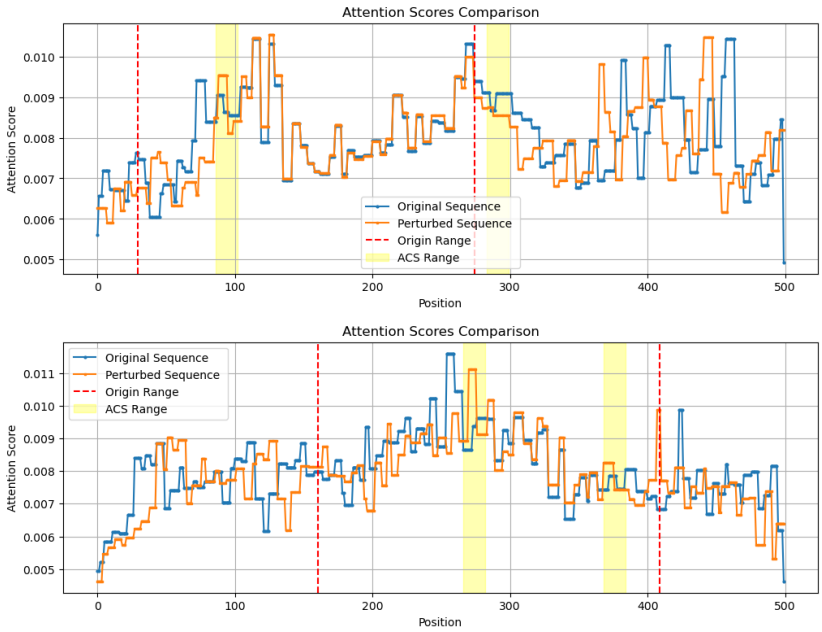


Figure 5.10: Attention score comparison for two samples after ACS motif elimination, with no change in model prediction.

An alternative approach to perturbing samples is randomly shuffling the tokenized sequences. By feeding these shuffled samples into the fine-tuned model, we evaluated the impact of this perturbation on classification probability to assess the model's sensitivity to token order.

We applied this aproach to a model fine-tuned on the Rang-Neg dataset, making predictions on the test set for true positive (TP) samples. Notably, after shuffling, all perturbed samples were still correctly classified as positive instances. This suggests that DNABERT-2 relies primarily on the content of tokens rather than their order, indicating that token presence and composition are the key discriminative factors. Therefore, the model relies on weighting to individual tokens rather than capturing positional dependencies.

Importance of Length of Tokenized Sequences: DNABERT-2 employs byte-pair encoding (BPE) tokenization, resulting in tokens of varying lengths within its vocabulary. Consequently, although all sequences in our dataset are initially of uniform length, the tokenized input sequences become variable in length. To address this, the model appends a [PAD] token to the end of shorter sequences, ensuring consistent input lengths for the model. We sought to compare origin and non-origin samples in terms of length after tokenization, for Rand-Neg dataset. Because origin sequences contain relatively long motifs which likely present in the BPE vocabulary, we hypothesized origin sequences would have shorter tokenized input lengths on average compared to non-origin sequences. To visualize the true length distribution, we removed the [PAD] tokens and then plotted the distribution of tokenized sequence length for both class. As shown in the histogram in Figure 5.11, origin sequences tend to have

shorter tokenized lengths than non-origin sequences due to the presence of longer tokens in origin regions.

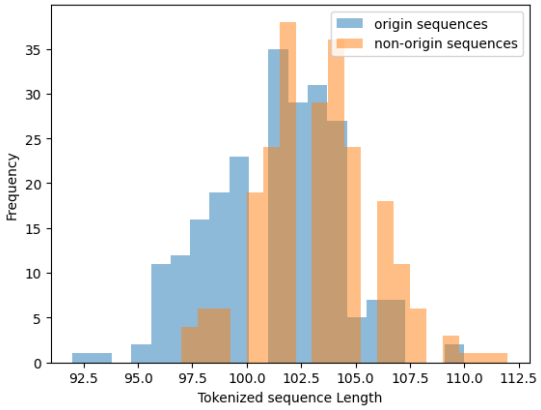


Figure 5.11: Distribution of tokenized sequence lengths for origin and non-origin sequences, by BPE tokenization.

Based on this histogram, when using only the tokenized input length as a predictor and setting a classification threshold of 103, we achieved an accuracy of 0.68. This suggests that input length itself may play a role in the model's ability to distinguish between the two classes. Moreover, input length can be reflected in attention maps, as shorter tokenized sequences require more [PAD] tokens at the end.

In conclusion, considering (1) DNABERT-2's comparable performance to DNABERT, (2) its strong diagonal attention scores in heatmaps and (3) insights provided by perturbation-base explanation; it can be suggested that its effectiveness and different attention maps are primarily due to its different tokenization method. BPE tokenization enables model to optimize token weighting within short-range interactions, improving classification despite reduced long-range attention. BPE compresses sequence information into fewer, longer tokens. Since longer tokens contain more meaningful sub-sequences, the model learns local patterns more effectively, boosting its ability to differentiate classes. This reduces the need for long-range attention because some tokens already encode richer subsequences. For example the longest token in our dataset has 32 bps length which was seen in an origin sequence.

While DNABERT captures both short-range and long-range dependencies, DNABERT-2's alternative tokenization approach appears to bias its learning towards short-range interactions by weighting tokens more effectively. This shift allows DNABERT-2 to achieve performance comparable to its predecessors.

Chapter 6

Discussion and Future Work

6.1 Conclusion

Our analysis of DNABERT and DNABERT-2 on the task of DNA replication origin prediction reveals distinct performance patterns and attention mechanism. While DNABERT achieves slightly better performance on the Rand-Neg dataset, DNABERT-2 remains competitive and even demonstrates slightly better performance on other datasets.

Based on our performance analysis across various datasets, we conclude that the tokenization approach plays a crucial role in both model performance and the way the attention mechanism operates in these models. Despite sharing the same BERT-style architecture, DNABERT-2's alternative tokenization strategy biases its learning toward short-range interactions, effectively optimizing token weighting. In contrast, DNABERT captures both short-range and long-range dependencies, as reflected in its attention maps, enabling it to model broader sequence relationships.

A key factor in the models' ability to distinguish replication origins is the presence of ARS consensus sequence (ACS) motifs. Our motif discovery pipeline, which extracts motifs from high-attention fragments of sequences, highlights DNABERT's ability to capture biologically relevant sequence patterns. The discovered motifs show a strong alignment with experimentally confirmed ACS patterns, indicating that these motifs play a crucial role in DNABERT's predictive power. Interestingly, while DNABERT-2 does not exhibit visually distinguishable high attention scores for ACS motifs, our perturbation-based explainability analysis confirms that eliminating these motifs significantly impacts its predictions, suggesting that ACS motifs also contribute to its decision-making process of this model.

Overall, while DNABERT excels in capturing both local and long-range dependencies, DNABERT-2 compensates through local-range dependencies by optimizing token weighting, achieving comparable performance despite. These insights provide valuable guidance for selecting LLMs for genomics applications, particularly in tasks requiring sequence pattern recognition and interpretability.

6.2 Future Work

As a potential direction for future work, one idea to explore is benchmarking our datasets on state-of-the-art foundation models or other existing genome language models. For instance, Evo-2 [41], a biological foundation model, presents an intriguing option as it offers interpretability techniques. Leveraging such models would allow us to evaluate the impact of different model architectures on replication origin prediction, potentially improving both predictive performance and interpretability in genomic studies.

Building on our work, where we applied DNABERT and DNABERT-2 for replication origin prediction in yeast, another idea can be extending these models to analyze human replication origins using datasets such as those presented by Petryk et al. [42]. The research provides a comprehensive genome-wide view of replication initiation and termination in the human genome, revealing how chromatin structure and transcription activity influence the replication program in a cell-type-specific manner. These findings offer valuable insights that can guide future applications of DNABERT and DNABERT-2 in DNA replication origin prediction beyond budding yeast, by fine-tuning the models on human replication origin datasets. Hence, one can evaluate model performance on datasets containing human initiation zones identified in [42]. Moreover, our explainability approaches can be applied to human replication origins to gain deeper insights into the sequence patterns and features influencing model predictions. In budding yeast, ACS motifs served as a gold standard, and our analysis showed that DNABERT effectively identified these motifs using its attention mechanism. However, in human replication origins, no well-defined consensus motifs are known. Therefore, a promising direction for future research is to apply our attention-based motif discovery approach to the human genome. If DNABERT achieves satisfactory performance when fine-tuned on human replication data, this method could help uncover potential sequence patterns that contribute to replication initiation in human cells.

Appendix A

Motifs Found For DNABERt-2

As mentioned in section 5.2.2, we tried to find motifs by adapting DNABERT find motif tool for DNABERT-2. We adjusted loser cut-off valuse for selecting high attention regions, as i) attention be greater than mean of attention within the sequence; and ii) contiguous regions must have a minimum length of L=6. Motifs with P-value< 0.05 are as follow:

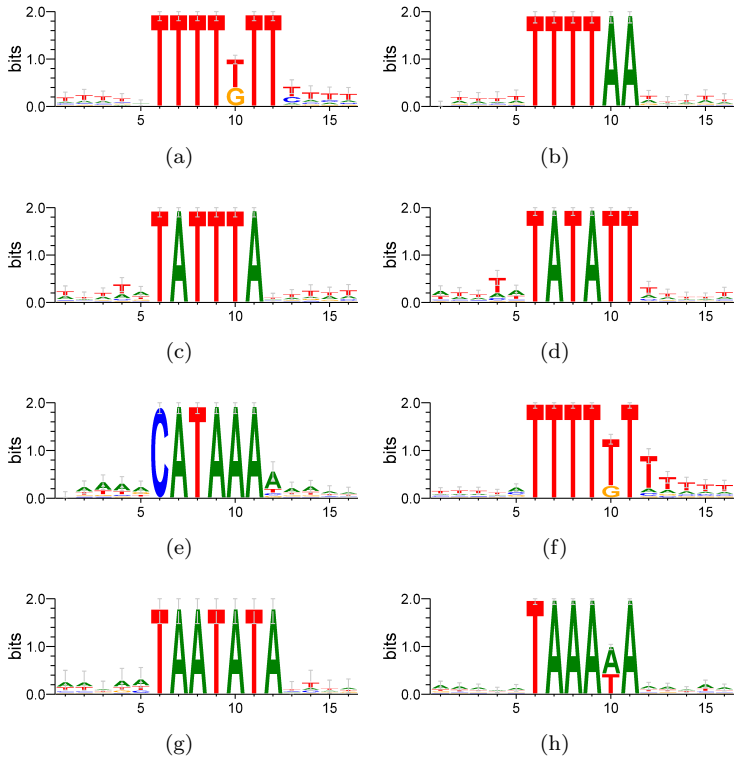


Figure A.1: Motifs found for DNABERT-2

Bibliography

[1] J. Boyle et al. *Molecular Biology of the Cell*, volume 36. Biochem. Mol. Biol. Educ., 5 edition, 2008.

[2] B. Ekundayo and F. Bleichert. Origins of dna replication. *PLoS Genetics*, 15(9):e1008320, 2019.

[3] F. Jacob, S. Brenner, and F. Cuzin. On the regulation of dna replication in bacteria. *Cold Spring Harbor Symposia on Quantitative Biology*, 28:329–348, 1963.

[4] D. Wang and F. Gao. Comprehensive analysis of replication origins in *Saccharomyces cerevisiae* genomes. *Frontiers in Microbiology*, 10:2122, 2019.

[5] C. S. K. Lee, M. Weiß, and S. Hamperl. Where and when to start: Regulating dna replication origin activity in eukaryotic genomes. *Nucleus*, 14(1), 2023.

[6] J. R. Broach, Y. Y. Li, J. Feldman, M. Jayaram, J. Abraham, K. A. Nasmyth, and J. B. Hicks. Localization and sequence analysis of yeast origins of dna replication. *Cold Spring Harbor Symposia on Quantitative Biology*, 47(Pt 2):1165–1173, 1983.

[7] Y. Marahrens and B. Stillman. A yeast chromosomal origin of dna replication defined by multiple functional elements. *Science (New York, N.Y.)*, 255(5046):817–823, 1992.

[8] J. F. Theis and C. S. Newlon. The ars309 chromosomal replicator of *Saccharomyces cerevisiae* depends on an exceptional ars consensus sequence. *Proceedings of the National Academy of Sciences of the United States of America*, 94(20):10786–10791, 1997.

[9] Cheuk C. Siow, Sian R. Nieduszynska, Carolin A. Müller, and Conrad A. Nieduszynski. Oridb, the dna replication origin database updated and extended. *Nucleic Acids Research*, 40(D1):D682–D686, 2012.

[10] J. Michael Cherry, Eurie L. Hong, Craig Amundsen, Rama Balakrishnan, Gail Binkley, Esther T. Chan, Karen R. Christie, Maria C. Costanzo, Selina S. Dwight, Stacia R. Engel, Dianna G. Fisk, Jodi E. Hirschman, Benjamin C. Hitz, Kalpana Karra, Cynthia J. Krieger, Stuart R. Miyasato, Rob S. Nash, Julie Park, Marek S. Skrzypek, Matt Simison, Shuai Weng, and Edith D. Wong. *Saccharomyces* genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, 40(D1):D700–D705, 2012.

[11] Feng Gao, Hao Luo, and Chun-Ting Zhang. Deori: a database of eukaryotic dna replication origins. *Bioinformatics*, 28(11):1551–1552, 2012.

[12] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, aug 2021.

[13] Zhihan Zhou, Yanrong Ji, Weijian Li, Pratik Dutta, Ramana V Davuluri, and Han Liu. DNABERT-2: Efficient foundation model and benchmark for multi-species genomes. In *The Twelfth International Conference on Learning Representations*, 2024.

[14] Wen-Chao Li, Zhe-Jin Zhong, Pan-Pan Zhu, En-Ze Deng, Hui Ding, Wei Chen, and Hao Lin. Sequence analysis of origins of replication in the Saccharomyces cerevisiae genomes. *Frontiers in Microbiology*, 5, 2014.

[15] A M Breier, S Chatterji, and N R Cozzarelli. Prediction of Saccharomyces cerevisiae replication origins. *Genome biology*, 5(4):R22, 2004.

[16] Wei Chen, Feng Pengmian, and Lin Hao. Prediction of replication origins by calculating dna structural properties. *FEBS Letters*, 586, 2012.

[17] Wen-Chao Li, En-Ze Deng, Hui Ding, Wei Chen, and Hao Lin. iori-pseknc: A predictor for identifying origin of replication with pseudo k-tuple nucleotide composition. *Chemometrics and Intelligent Laboratory Systems*, 141:100–106, 2015.

[18] Wei Chen, Tian-Yu Lei, Dian-Chuan Jin, Hao Lin, and Kuo-Chen Chou. Pseknc: A flexible web server for generating pseudo k-tuple nucleotide composition. *Analytical Biochemistry*, 456:53–60, 2014.

[19] X Xiao, HX Ye, Z Liu, JH Jia, and KC Chou. iros-gpseknc: Predicting replication origin sites in dna by incorporating dinucleotide position-specific propensity into general pseudo nucleotide composition. *Oncotarget*, 7(23):34180–34189, jun 2016.

[20] Bin Liu, Fan Weng, De-Shuang Huang, and Kuo-Chen Chou. iro-3wpseknc: identify dna replication origins by three-window-based pseknc. *Bioinformatics*, 34(18):3086–3093, 2018.

[21] Bin Liu, Shengyu Chen, Ke Yan, and Fan Weng. iro-psekgcc: Identify dna replication origins based on pseudo k-tuple gc composition. *Frontiers in Genetics*, 10, 2019.

[22] Fu-Ying Dao, Hao Lv, Fang Wang, Chao-Qin Feng, Hui Ding, Wei Chen, and Hao Lin. Identify origin of replication in *Saccharomyces cerevisiae* using two-step feature selection technique. *Bioinformatics*, 35(12):2075–2083, 2019.

[23] Hao Lin, En-Ze Deng, Hui Ding, Wei Chen, and Kuo-Chen Chou. ipro54-pseknc: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition. *Nucleic Acids Research*, 42(21):12961–12972, 2014.

[24] P. A. Mundra and J. C. Rajapakse. Svm-rfe with mrmr filter for gene selection. *IEEE Transactions on NanoBioscience*, 9(1):31–37, March 2010.

[25] Vinod Kumar Singh, Vipin Kumar, and Annangarachari Krishnamachari. Prediction of replication sites in *Saccharomyces cerevisiae* genome using dna segment properties: Multi-view ensemble learning (mel) approach. *Biosystems*, 163:59–69, 2018.

[26] Duyen Thi Do and Nguyen Quoc Khanh Le. Using extreme gradient boosting to identify origin of replication in *Saccharomyces cerevisiae* via hybrid features. *Genomics*, 112(3):2445–2451, 2020.

[27] Balachandran Manavalan, Shaherin Basith, Tae Hwan Shin, and Gwang Lee. Computational prediction of species-specific yeast dna replication origin via iterative feature representation. *Briefings in Bioinformatics*, 22(4), 2021.

[28] F. Wu, R. Yang, C. Zhang, et al. A deep learning framework combined with word embedding to identify dna replication origins. *Scientific Reports*, 11:844, 2021.

[29] Mahwish Shahid, Maham Ilyas, Waqar Hussain, and Yaser Daanial Khan. Ori-deep: improving the accuracy for predicting origin of replication sites by using a blend of features and long short-term memory network. *Briefings in Bioinformatics*, 23(2), 2022.

[30] Zhen-Ning Yin, Fei-Liao Lai, and Feng Gao. Unveiling human origins of replication using deep learning: accurate prediction and comprehensive analysis. *Briefings in Bioinformatics*, 25(1), 2024.

[31] Conrad A. Nieduszynski, Shin-ichiro Hiraga, Pinar Ak, Craig J. Benham, and Anne D. Donaldson. Oridb: a dna replication origin database. *Nucleic Acids Research*, 35(Database issue):D40–D46, 2007.

[32] Sven Heinz, Christopher Benner, Nathanael Spann, Eric Bertolino, Yin C. Lin, Peter Laslo, Jason X. Cheng, Cornelis Murre, Harinder Singh, and Christopher K. Glass. Simple combinations of lineage-determining transcription factors prime *cis*-regulatory elements required for macrophage and b cell identities. *Molecular Cell*, 38(4):576–589, 2010.

[33] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *CoRR*, 2018.

[34] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Ilya Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, 2017.

[35] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany, August 2016. Association for Computational Linguistics.

[36] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *arXiv preprint arXiv:2108.12409*, 2021.

[37] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher R. Flashattention: Fast and memory-efficient exact attention with io-awareness, 2022.

[38] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *CoRR*, abs/2106.09685, 2021.

[39] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Trans. Intell. Syst. Technol.*, 15(2), 2024.

[40] Timothy L. Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. In *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pages 28–36, Menlo Park, California, 1994. AAAI Press.

[41] Garyk Brixi, Matthew G. Durrant, Jerome Ku, Michael Poli, Greg Brockman, Daniel Chang, Gabriel A. Gonzalez, Samuel H. King, David B. Li, Aditi T. Merchant, Mohsen Naghipourfar, Eric Nguyen, Chiara Ricci-Tam, David W. Romero, Gwanggyu Sun, Ali Taghibakshi, Anton Vorontsov, Brandon Yang, Myra Deng, Liv Gorton, Nam Nguyen, Nicholas K. Wang, Etowah Adams, Stephen A. Baccus, Steven Dillmann, Stefano Ermon, Daniel Guo, Rajesh Ilango, Ken Janik, Amy X. Lu, Reshma Mehta, Mohammad R.K. Mofrad, Madelena Y. Ng, Jaspreet Pannu, Christopher Ré, Jonathan C. Schmok, John St. John, Jeremy Sullivan, Kevin Zhu, Greg Zynda, Daniel Balsam, Patrick Collison, Anthony B. Costa, Tina Hernandez-Boussard, Eric Ho, Ming-Yu Liu, Thomas McGrath, Kimberly Powell, Dave P. Burke, Hani Goodarzi, Patrick D. Hsu, and Brian L. Hie. Genome modeling and design across all domains of life with evo 2. *bioRxiv*, 2025.

[42] Nicolas Petryk, Mehdi Kahli, Yvan d'Aubenton Carafa, et al. Replication landscape of the human genome. *Nature Communications*, 7:10208, 2016.