

FEDERAL UNIVERSITY OF TECHNOLOGY - PARANÁ
CAMPUS DOIS VIZINHOS
BIOTECHNOLOGY AND BIOPROCESS ENGINEERING COURSE

Israel Yago Pereira

**Automatic gene annotation with Artificial
Intelligence: Binary classification between
enzymes and non-enzymes**

Dois vizinhos, Paraná

2024

Israel Yago Pereira

**Automatic gene annotation with Artificial Intelligence:
Binary classification between enzymes and non-enzymes**

Conclusion work presented to the course of
Biotechnology and Bioprocess Engineering as
a partial requirement for obtaining the title of
Biotechnology and Bioprocess Engineer.

Federal University of Technology - Paraná

Supervisor: Marlon Marcon

Co-supervisor: Naiana Cristine Gabiatti

Dois vizinhos, Paraná

2024

ACKNOWLEDGEMENTS

Many thanks to these modern Large Language Models like ChatGPT and their influence.

*"From the moment I understood the weakness of my flesh, it disgusted me.
I craved the strength and certainty of steel.
I aspired to the purity of the Blessed Machine.
Your kind cling to your flesh, as though it will not decay and fail you.
One day the crude biomass you call the temple will wither,
and you will beg my kind to save you.
But I am already saved, for the Machine is immortal...
Even in death I serve the Omnissiah."
(Warhammer 40.000)*

RESUMO

A anotação genômica, um passo pivotal na gênica, implica descobrir elementos funcionais, tais como genes e componentes regulatórios, dentro das sequências de DNA. Esse processo é crucial para compreender processos biológicos e apontar mutações relacionadas às doenças. A integração de ferramentas de alto desempenho de sequenciamento de DNA e ferramentas computacionais tem revolucionado a anotação gênica, garantindo elevada acurácia através da integração de dados. A anotação genética manual, envolvendo a identificação e anotação de diversos elementos genômicos, é muito trabalhosa, tornando métodos tradicionais desafiadores devido à intrínseca natureza dos dados genômicos e diversidade de espécies, além do contínuo influxo de novas informações gênicas com diferentes práticas de anotação entre grupos de pesquisadores, o que dificulta ainda mais o problema. Nesse trabalho, pretende-se criar um *dataset* de sequências de amino ácidos com a classe binária de enzima e não-enzima e um modelo para classificar as sequências em enzimas e não-enzimas, eliminando alguns dos problemas atuais do processo de anotação gênica. O *dataset* será compilado a partir do *UnitProt Consortium*, contendo tanto sequências de amino ácidos de enzimas e não-enzimas representadas no formato padrão FASTA. O núcleo da arquitetura será adaptado do seguimento codificador do Transformador, reconhecido por sua capacidade de capturar dependências intrínsecas dentre os dados sequenciais. Cada amino ácido na sequência será tratado analogamente como um *token* e a arquitetura adaptada excluirá a componente decodificadora por ser desnecessária na formulação do problema. As considerações do tamanho do modelo serão baseadas tanto no orçamento computacional quanto na quantidade de *tokens*.

Palavras-chave: Inteligência Artificial. Transformador. Enzima.

ABSTRACT

Genomic annotation, a pivotal step in genomics, entails uncovering functional elements such as genes and regulatory components within DNA sequences. This process is crucial for comprehending biological processes and pinpointing disease-related mutations. Integrating high-throughput DNA sequencing and computational tools has revolutionized genetic annotation, ensuring heightened accuracy through data integration. Manual genetic annotation, involving the identification and annotation of diverse genomic elements, is labor-intensive, making traditional methods challenging due to the intricate nature of genomic data, species diversity, and the continual influx of new genomic information with different annotation practices among research groups also makes the problem harder. In this work, we intend to create a dataset of amino acid sequences with the binary class of enzymes and non-enzymes and a model for classifying enzymes and non-enzymes sequences, eliminating some of these current problems of the genomic annotation pipeline. The dataset will be compiled from The UniProt Consortium, encompassing both enzyme and non-enzyme amino acid sequences represented in standard FASTA format. The core of the model architecture will be adapted from the Transformer's encoder segment, renowned for its ability to capture intricate dependencies within sequential data. Each amino acid in the sequence will be treated as an analogous token, and the adapted architecture excludes the decoder component as it is unnecessary for the problem formulation. Model size considerations are based on both computing budget and token quantity.

Keywords: Artificial Intelligence. Transformer. Enzyme.

CONTENTS

1	INTRODUCTION	9
1.1	OBJECTIVES	10
1.1.1	General objectives	10
1.1.2	Specific objectives	10
2	THEORETICAL FRAMEWORK	11
2.1	BIOLOGICAL FRAMEWORK	11
2.1.1	Amino acids	11
2.1.2	Enzymes	12
2.2	GENETIC ANNOTATION	14
2.3	ARTIFICIAL INTELLIGENCE (AI)	16
2.3.1	Convolutional Neural Networks (CNNs)	16
2.3.2	Recurrent Neural Networks (RNNs)	16
2.3.3	Diffusion models	17
2.3.4	Consistency models	18
2.3.5	Transformer	18
2.4	BIOINFORMATICS AND MACHINE LEARNING	19
2.4.1	Bioinformatics	19
2.4.2	Alphafold2	19
2.4.3	Bidirectional Encoder Representations from Transformers (BERT)	20
2.4.4	GenomicBERT	21
2.4.5	BERT-Promoter	21
2.4.6	Related work	21
3	MATERIALS AND METHODS	22
3.1	TOOLS	22
3.2	DATASET SELECTION AND PREPROCESSING	22
3.3	TRANSFORMER ENCODER ADAPTATION	22
3.4	MODEL SIZE	22

3.5	MODEL TRAINING	23
3.6	EVALUATION AND VALIDATION	25
3.7	SCHEDULE	25
4	EXPECTED RESULTS	26
	 BIBLIOGRAPHY	 27
	 ANNEX	 30
	ANNEX A – EXAMPLE OF A FASTA FILE	31

1 INTRODUCTION

Enzymes are biological molecules that act as catalysts in living organisms without suffering any overall change (PALMER; BONNER, 2011; COPELAND, 2023). They play a vital role in facilitating and accelerating biochemical reactions by lowering the activation energy required for these reactions to occur. Enzymes are essential for various biological processes, including metabolism, digestion, DNA replication, and cell signaling. Based on the type of reaction the enzyme catalyzes, it can be classified into one of (PALMER; BONNER, 2011): Oxidoreductases, Transferases, Hydrolases, Lyases, Isomerases, and Ligases.

Artificial Intelligence techniques are capable of capturing intricate hierarchical features from complex input data, allowing them to be helpful for a diverse range of tasks, including image analysis tasks, image generation, machine translation, text generation, sentiment analysis, and audio generation (ZHIQIANG; JUN, 2017; WANG; WAN; WAN, 2020; VASWANI et al., 2023; LAKEW; CETTOLO; FEDERICO, 2018; KREUK et al., 2023; TOUVRON et al., 2023; SHU et al., 2022; DOSOVITSKIY et al., 2021)

The Transformer architecture, made for Natural Language Processing (NLP), is remarkably good at paying attention to each input and how it relates to all other inputs, the self-attention mechanism. In the biological field, we have the use of architectures that use a Transformer deep down, like GenomicBERT (CHEN et al., 2023), BERT-promoter (LE et al., 2022), ProteinBERT (BRANDES et al., 2022), and the famous AlphaFold2 (JUMPER et al., 2021), which although does not use a Transformer, also has an attention mechanism.

Genetic annotation is a crucial process in genomics, involving identifying functional elements in DNA sequences. This process includes genes, regulatory elements, and more, providing insights into gene functions and regulatory mechanisms. It is vital in basic research and medical genomics to understand fundamental biological processes and identify disease-causing mutations. Advances in high-throughput DNA sequencing have transformed genetic annotation, with computational tools and data integration enhancing accuracy. Genetic annotation is a sophisticated process combining experimental data and computational tools to decode the functional elements within a genome (SOH; GORDON; SENSEN, 2012; RUST; MONGIN; BIRNEY, 2002).

Manual genetic annotation involves identifying and annotating genes, regulatory elements, and other functional elements within a genome and is labor-intensive and time-consuming. The complexity of genomic data, the diversity of species, and the rapid accumulation of new genomic information pose significant challenges to traditional annotation methods. Additionally, inconsistencies and variations in annotation practices among different research groups can lead to discrepancies in the interpretation of genomic data.

1.1 OBJECTIVES

1.1.1 General objectives

In this work, we intend **to create a dataset of amino acid sequences tied with the binary class of enzymes and non-enzymes and an automated tool for annotation of enzymes and non-enzymes amino acid sequences**, allowing for future work to annotate the actual function of such enzymes, growing our understanding of the genetic material.

1.1.2 Specific objectives

We intend to collect amino acid sequences from UniProt (CONSORTIUM, 2022), organized by the information about whether or not the sequence refers to an enzyme.

Then, we plan to utilize the encoder part of the Transformer architecture, with a dictionary of 20 Amino acids and 128 input tokens, with a final binary classifier for enzymes and non-enzymes based solely on amino acid sequences collected earlier. All other parameters will be the same as used in the original Transformer architecture (VASWANI et al., 2023).

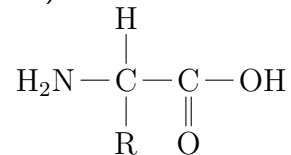
2 THEORETICAL FRAMEWORK

2.1 BIOLOGICAL FRAMEWORK

2.1.1 Amino acids

Enzymes are generally proteins, and proteins are made from amino acids joined in a very specific series, determined by the genetic code. Given that there are multiples sequences, each protein acquires unique properties.

Amino acids are classified as organic compounds featuring both an amino group (-NH_2 or $>\text{NH}$) and a carboxyl group (-COOH). Consequently, they exhibit characteristics that encompass traits of both bases and acids. The general formula for an amino acid with a side chain R is (PALMER; BONNER, 2011):



And the following side chains are described on the literature:

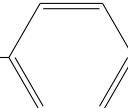
Non-polar side chains:

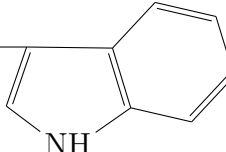
Alanine (Ala) (A): $\text{R}-\text{CH}_3$

Valine (Val) (V): $\text{R}-\underset{\text{CH}_3}{\text{CH}}-\text{CH}_3$

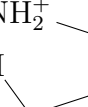
Leucine (Leu) (L): $\text{R}-\text{CH}_2-\underset{\text{CH}_3}{\text{CH}}-\text{CH}_3$

Isoleucine (Ile) (I): $\text{R}-\underset{\text{CH}_3}{\text{CH}}-\text{CH}_2-\text{CH}_3$

Phenylalanine (Phe) (F): $\text{R}-\text{CH}_2-$ 

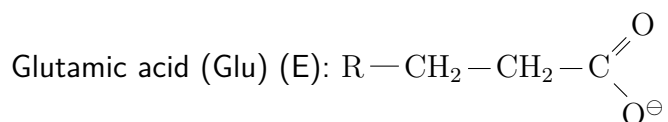
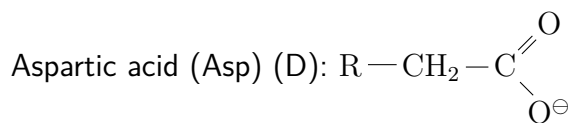
Tryptophan (Trp) (W): $\text{R}-\text{CH}_2-$ 

Methionine (Met) (M): $\text{R}-\text{CH}_2-\text{CH}_2-\text{S}-\text{CH}_3$

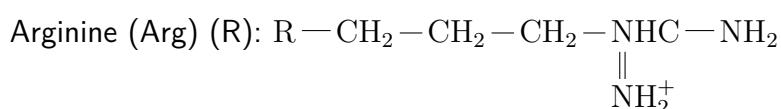
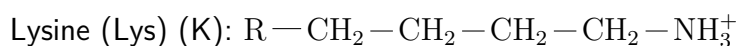
Proline (Pro) (P): $\text{CO}_2^- - \underset{\text{NH}_2^+}{\text{CH}}$ 

Polar side chains:

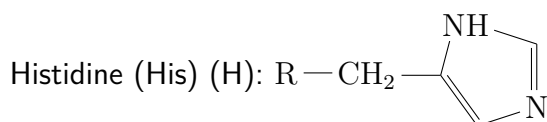
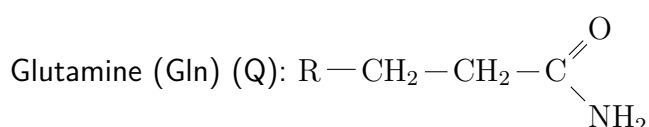
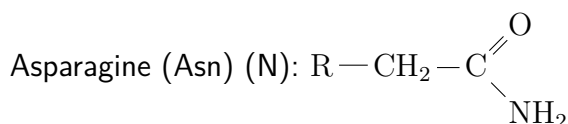
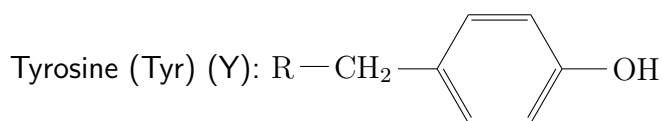
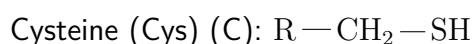
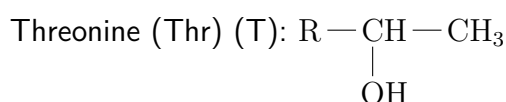
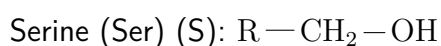
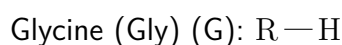
Negative charge at pH 7:



Positive charge at pH 7:



Uncharged at pH 7:



Note that Proline is not an α -amino acid, but rather it is an imino acid, deviating for the R common structure. Generally speaking, the proteins are almost exclusively of L-amino acids, with a possible explanation the specificity of enzymes, by chance, selected L- isomers instead of D-. Most naturally occurring amino acids are in the L- isomeric form (PALMER; BONNER, 2011).

2.1.2 Enzymes

Enzymes increase the rate of chemical reactions without suffering in the process, thus, they are biological catalysts. Some enzymes require the use of a non-protein component, a

cofactor.

Enzymes are typically classified into several major categories based on the types of reactions they catalyze (PALMER; BONNER, 2011):

Oxidoreductases: These enzymes catalyze oxidation-reduction reactions, involving the transfer of hydrogen atoms, oxygen atoms, or electrons between substrates. Examples include dehydrogenases, oxidases, and reductases.

Transferases: Transferases aid the transfer of functional groups from one molecule to another. Examples include kinases, methyltransferases, and transaminases.

Hydrolases: These enzymes catalyze hydrolysis reactions, where water molecules are used to break chemical bonds. Lipases, proteases, and nucleases are examples of hydrolases.

Lyases: Lyases catalyze the removal or addition of groups from molecules without using water. Decarboxylases and deaminases are examples of lyases.

Isomerases: Isomerases convert molecules between different structural isomers (e.g., cis to trans) or rearrange atoms within a molecule. Examples include racemases and mutases.

Ligases: Ligases catalyze the joining of two molecules, often using energy from ATP. DNA ligase, which joins DNA strands, is an example of a ligase.

Amino acids are the building block of enzymes, upon which the sequence of the enzyme/protein is very specific. Such sequence is determined by the DNA of the living organism. Simple proteins are proteins made only by a sequence of amino acids, while conjugated proteins have more material bounded in these amino acids. Enzymes too, can be either simple or conjugated proteins.

The amount of properties shown by proteins in general can be attributed by the variety of side chain characteristics that mixture of the amino acids can produce. The structure of a protein can be primary, secondary, tertiary or quaternary.

Primary structures are composed of a chain of polypeptide, which in turn, a polypeptide chain is the residues of amino acids that have undergone a condensation reaction. Given that a molecule can rotate freely in a single covalent bond, polypeptide chains can have basically an unlimited number of arrangements in space, contributing to their unique properties. Secondary chains are the repeating patterns of part of a backbone of the polypeptide chain stabilised by hydrogen bonding. Some amino acids cannot be stabilised in these secondary structures, thus, they disrupt the arrangement. Although there could be an greater number of different structures resulting from this, each polypeptide chain happens to have a specific three-dimensional structure. This structure is called the tertiary structure, stabilised by amino acids who bond themselves when in close proximity.

When multiple, identical or not, polypeptide chains are linked together forming the protein and their interactions, the whole structure is called the quaternary structure (PALMER;

BONNER, 2011).

Even though we do have the possibility that these chains are coded from different parts of the DNA, and only a sequence of amino acids may not tell directly everything about the protein, in this work we make the consideration that given the sequence, even if it's incomplete, it is enough to make reasonable predictions about the function of the protein. In our case, to classify the protein in enzyme and non-enzyme.

2.2 GENETIC ANNOTATION

Genetic annotation is a process within the realm of genomics that involves deciphering and identifying various functional elements within a DNA sequence. In essence, genetic annotation elucidates the underlying significance and role of specific regions within the genome. These annotations provide critical insights into the functions, structures, and regulatory mechanisms of genes, non-coding regions, and other essential genomic elements (SOH; GORDON; SENSEN, 2012).

The process of genetic annotation encompasses the categorisation and labeling of distinct features within a DNA sequence, each of which contributes to the overall understanding of the genetic blueprint. These features include protein-coding genes, non-coding RNA genes, transcription factor binding sites, regulatory elements, promoters, enhancers, exons, introns, and splice sites, among others. The process extends to elucidating the potential effects of genetic variants, mutations, and polymorphisms on protein function, gene expression, and disease susceptibility.

Genetic annotation plays a role in both basic research and applied fields. In basic research, it aids in unraveling the intricate mechanisms that govern gene expression, regulation, and interaction, thus contributing to the comprehension of fundamental biological processes. In applied fields, genetic annotation holds immense significance in the context of medical genomics, as it assists in identifying disease-causing mutations, understanding the genetic basis of hereditary disorders, and tailoring personalized treatment strategies.

Over time, the methods employed for genetic annotation have evolved in tandem with technological advancements. Traditional methods relied on experimental techniques to identify genes and functional elements. However, the advent of high-throughput DNA sequencing technologies has revolutionized genetic annotation by enabling the rapid and cost-effective acquisition of vast genomic data. As a result, computational tools and algorithms have become indispensable for processing and interpreting the large amount of genomic information (RUST; MONGIN; BIRNEY, 2002).

Modern genetic annotation often involves the integration of diverse data sources, including DNA sequence information, epigenomic data, transcriptomic data, and comparative genomics data. This multifaceted approach enhances the accuracy and reliability of annotations,

providing a holistic view of genomic function.

Genetic annotation in the present day is a sophisticated and multifaceted process that combines experimental data, computational algorithms, and extensive databases to decode the functional elements within a genome. The process involves identifying genes, regulatory elements, and other functional regions, as well as characterizing their roles and interactions. Here's an overview about genetic annotation (SOH; GORDON; SENSEN, 2012):

DNA Sequencing: Modern genetic annotation heavily relies on high-throughput DNA sequencing technologies, which can rapidly generate massive amounts of genomic data.

Epigenomics: Epigenetic modifications, such as DNA methylation and histone modifications, provide crucial information about gene regulation.

***Ab Initio* Prediction:** Computational algorithms scan DNA sequences for characteristic features like start codons, stop codons, and open reading frames to predict protein-coding genes.

Homology-Based Prediction: Comparative genomics involves comparing the target genome with well-annotated reference genomes to identify conserved gene sequences.

Small RNA: Specialized algorithms identify small non-coding RNAs, like microRNAs and small nucleolar RNAs, which play regulatory roles.

Long Non-Coding RNA (lncRNA): Machine learning approaches predict lncRNAs based on sequence and structural features.

Promoters and Enhancers: Computational methods identify promoter regions, enhancers, and transcription factor binding sites based on DNA sequence motifs and epigenetic marks.

Chromatin Accessibility Data: Techniques like ATAC-seq and DNase-seq provide insights into regions of open chromatin, indicative of potential regulatory elements (TSOMPANA; BUCK, 2014).

Protein Function Prediction: Sequence homology, protein domain analysis, and structural modeling (JUMPER et al., 2021) aid in predicting protein functions.

Variant Annotation: Genetic variants (mutations) are annotated to determine potential effects on gene function, such as missense mutations or splice site disruptions.

Databases: Publicly available databases like GenBank (BENSON et al., 2012), Ensembl (CUNNINGHAM et al., 2022), and UCSC Genome Browser (DRESZER et al., 2012) provide comprehensive annotations and data resources for various species.

Integration: Integrative tools combine multiple data sources to refine annotations and provide a holistic view of gene regulation.

Functional Assays: Experimental techniques like CRISPR-Cas9 gene editing and RNA interference validate the roles of predicted genes and regulatory elements.

Expression Profiling: Transcriptomics, such as RNA-seq, provides insights into gene expression patterns under different conditions.

Genome Browsers: Interactive genome browsers allow researchers to visualize annotated features in the context of the entire genome.

Pathway and Network Analysis: Software tools help interpret annotations by placing genes and regulatory elements in the context of biological pathways and networks.

Although our understanding of genomics deepens and technology continues to advance, genetic annotation methodologies need to become even more sophisticated and automatic, leading to increasingly accurate and detailed annotations of genomes given large-scale data produced in the field (RUST; MONGIN; BIRNEY, 2002).

2.3 ARTIFICIAL INTELLIGENCE (AI)

2.3.1 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks represent a architecture that has significantly revolutionized the field of computer vision and pattern recognition. Inspired by the visual processing mechanisms of the human brain, CNNs are adept at capturing intricate hierarchical features from complex input data, making them particularly well-suited for image analysis tasks (ZHIQIANG; JUN, 2017).

At its core, a CNN is structured as a multi-layered neural network, composed of convolutional layers, pooling layers, and fully connected layers. The convolutional layers are the hallmark of CNNs, employing learnable filters to convolve across input data, extracting localized features that are progressively refined through subsequent layers. This process effectively endows CNNs with the ability to automatically learn distinctive and discriminative features, obviating the need for manual feature engineering.

The hierarchical arrangement of layers empowers CNNs to capture features at varying levels of abstraction. Pooling layers, interspersed between convolutional layers, serve to down-sample the spatial dimensions of feature maps, enhancing computational efficiency and inducing a degree of translation invariance. The final fully connected layers consolidate these abstract features into class probabilities, enabling accurate object classification.

2.3.2 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks stand as a architectural innovation that has significantly reshaped the landscape of sequential data analysis, language modeling, and time-series prediction. Rooted in the aspiration to model temporal dependencies within data, RNNs have emerged as a powerful framework for processing sequences of variable lengths, holding considerable promise for a spectrum of applications. This exposition elucidates the foundational principles

and operational mechanics of RNNs, spotlighting their profound impact on tasks like natural language processing, speech recognition, and sequential data generation.

At its core, an RNN comprises a network structure that encompasses feedback loops, allowing information to persist across time steps, thus facilitating the modeling of sequential relationships. This inherent memory-like mechanism addresses the limitations of traditional feedforward neural networks by endowing RNNs with the ability to capture dependencies and contextual nuances present in sequential data (SCHUSTER; PALIWAL, 1997).

The crux of the RNN's operation resides in its recurrent connections, which enable the network to process each time step while considering both the current input and the previously processed input. This recurrent computation mechanism embodies a form of dynamic memory that enables the network to maintain a contextual understanding of the sequence, enabling applications ranging from sentiment analysis in text to melody generation in music.

2.3.3 Diffusion models

Diffusion Models have emerged as a cutting-edge paradigm that showcases remarkable potential in crafting highly realistic and intricate visual content. Representing a pioneering advancement in the domain of generative models, Diffusion Models have revolutionized the art of generating images by effectively modeling the intricate interplay between data and noise. This exposition delves into the foundational principles and operational intricacies of Diffusion Models, elucidating their profound impact on diverse applications, including image synthesis, style transfer, and anomaly detection.

At its essence, a Diffusion Model stands as an ingenious departure from traditional generative models by introducing a dynamic diffusion process into the image generation process. This process hinges on iteratively introducing carefully controlled levels of noise into an image, thereby facilitating the gradual transformation of the image from a latent noise state to a coherent and detailed representation. By modeling this diffusion process, the model inherently grasps the nuanced distribution of pixel intensities and their intricate relationships, culminating in images that possess an unparalleled level of realism and diversity.

Central to the functioning of Diffusion Models is the realization that the diffusion process mimics a series of progressive steps, each contributing to the refinement of the generated image. These steps are intricately orchestrated through a network architecture that encapsulates both a generator and a diffusion process simulator. The generator constructs an initial noisy image, while the diffusion simulator navigates the intricate trajectory of noise incorporation, allowing the image to gradually evolve while maintaining its fidelity to the underlying data distribution.

From high-resolution image synthesis and image-to-image translation to the synthesis of novel visual concepts, Diffusion Models have seamlessly integrated themselves into the fabric of contemporary AI applications.

2.3.4 Consistency models

Consistency models signify a groundbreaking evolution in the realm of generative models, addressing a fundamental limitation observed in the progression of diffusion models that have notably propelled the domains of image, audio, and video generation. While diffusion models have ushered in unprecedented realism and diversity, their reliance on iterative sampling introduces a perceptible constraint on the speed of generation. In response to this constraint, the authors from OpenAI have developed a paradigm-shifting innovation: consistency models. This novel category of models redefines the landscape of generative methods by directly mapping noise to data, thereby bypassing the iterative process and enabling swift one-step generation (SONG et al., 2023).

2.3.5 Transformer

The Transformer architecture represents a seismic shift in the realm of Artificial Intelligence, fundamentally redefining the landscape of natural language processing and sequence-to-sequence tasks. Emerging as a pivotal innovation, the Transformer has revolutionized the mechanisms underlying language understanding, translation, and contextual reasoning. This elucidation delves into the foundational principles and operational intricacies of the Transformer, illuminating its profound impact on diverse applications, including machine translation, text generation, sentiment analysis, audio generation and even image applications (WANG; WAN; WAN, 2020; VASWANI et al., 2023; LAKEW; CETTOLO; FEDERICO, 2018; KREUK et al., 2023; TOUVRON et al., 2023; SHU et al., 2022; DOSOVITSKIY et al., 2021).

At its core, the Transformer architecture is an intricate departure from traditional sequential models, marked by a self-attention mechanism that endows it with a holistic understanding of contextual relationships within a sequence. Unlike conventional recurrent or convolutional architectures, the Transformer eschews sequential processing, allowing for parallelism and eliminating the sequential bottlenecks that previously impeded scalability.

Central to the Transformer's efficacy is the mechanism of self-attention. This mechanism enables the architecture to weigh the significance of each token in relation to all other tokens within the same sequence, thereby fostering a comprehensive understanding of context. By capturing both local and global dependencies, the Transformer attains a remarkable capacity to handle long-range relationships and nuances present in natural language (VASWANI et al., 2023).

The Transformer has two core components: the encoder and decoder stacks. The encoder stack meticulously encodes input sequences into context-rich representations through multi-headed self-attention and feedforward neural networks. The decoder stack subsequently generates target sequences, leveraging an additional attention mechanism to incorporate information from the input and previously generated tokens.

Of pivotal significance is the concept of "positional encoding," which imparts the Transformer with an inherent awareness of token positions within a sequence. This mechanism obviates the need for sequential processing, allowing the architecture to distinguish tokens based on their positional attributes.

The adaptability of the Transformer to a plethora of applications, from language translation (BARRAULT et al., 2023) and text summarization to question answering (TOUVRON et al., 2023), sentiment analysis and audio generation (CHOI et al., 2020; KREUK et al., 2023), the Transformer has seamlessly integrated itself into the fabric of contemporary AI-driven language tasks. Its remarkable capabilities extend to accommodating diverse sequence lengths and languages, making it an indomitable force in multilingual and cross-lingual contexts (LAKEW; CETTOLO; FEDERICO, 2018).

By embracing self-attention, parallelism, and positional encoding, the Transformer has transcended conventional sequential limitations, culminating in an architecture that epitomizes the pinnacle of contextual understanding and sequence manipulation. Through its foundational principles and architectural innovation, the Transformer has firmly established itself as a bedrock of modern natural language understanding and generation.

2.4 BIOINFORMATICS AND MACHINE LEARNING

2.4.1 Bioinformatics

Bioinformatics can be described as the collection, storage, analysis, interpretation of biological data (PALMER; BONNER, 2011). One the most important database for DNA sequence information is from European Molecular Biology Laboratory (EMBL), and for annotated protein sequence, the SWISS-PROT.

2.4.2 Alphafold2

AlphaFold2 (JUMPER et al., 2021) is a groundbreaking deep learning-based model developed by DeepMind, that addresses the challenging problem of protein structure prediction. It represents a remarkable advancement in the field of structural biology by accurately predicting the three-dimensional structures of proteins, which is crucial for understanding their functions and interactions.

Deciphering the 3D structure of a protein from its amino acid sequence is known as the protein folding problem, and it has been a longstanding challenge in computational biology. Accurate protein structure prediction holds immense potential for various applications, including drug discovery, understanding diseases, and designing novel enzymes. (DILL; MACCALLUM, 2012)

AlphaFold2 employs a deep learning architecture, with similar principles used in natural language processing, to predict protein structures with remarkable accuracy. It takes the

sequence of amino acids that make up a protein as input and outputs a 3D atomic-level model of the protein's structure. The model's training involves learning from a vast dataset of protein structures, and the learned representations are used to predict the spatial arrangement of atoms in the protein.

What sets AlphaFold2 apart is its ability to predict protein structures with near-experimental accuracy. In the 14th edition of the Critical Assessment of Structure Prediction (CASP14) competition, AlphaFold2 significantly outperformed other methods in predicting the 3D structures of proteins. Its predictions were comparable to high-resolution experimental techniques like X-ray crystallography and cryo-electron microscopy (ROBERTSON et al., 2021; JONES; THORNTON, 2022).

AlphaFold 2's success can be attributed to its sophisticated architecture that combines deep learning, attention mechanisms, and a thorough understanding of protein biophysics. Its ability to capture long-range dependencies and contextual information within protein sequences helps it model the intricate folding patterns more accurately than previous methods.

2.4.3 Bidirectional Encoder Representations from Transformers (BERT)

The Bidirectional Encoder Representations from Transformers (BERT) architecture has emerged as a pivotal milestone in the current artificial intelligence landscape, revolutionizing natural language processing (NLP) tasks. BERT is a deep learning model that leverages the Transformer architecture's self-attention mechanisms to comprehend and generate contextualized representations of words within sentences (DEVLIN et al., 2019).

Unlike traditional language models that process text in a unidirectional or fixed context manner, BERT introduces bidirectional context by training on both left and right contexts of a word. This enables it to capture the rich interdependencies and nuances present in natural language, leading to more accurate understanding and contextual representation.

BERT's training process involves pretraining on massive amounts of text data to learn general language features. During this phase, it learns to predict missing words in sentences given the surrounding context. This unsupervised pretraining endows BERT with a robust understanding of language patterns and relationships. Subsequently, fine-tuning on task-specific datasets adapts the model to perform a wide range of natural language understanding tasks, such as text classification, question answering, and named entity recognition (RAFFEL et al., 2019).

The profound impact of BERT stems from its ability to generate high-quality contextualized word embeddings, which capture not only syntactic but also semantic relationships within text. This allows BERT to outperform previous language models on various benchmarks, even with minimal task-specific data. BERT's success has paved the way for subsequent advancements in natural language understanding and downstream applications across industries.

In essence, BERT's bidirectional context modeling, pretraining-finetuning framework, and contextualized embeddings have propelled it to a position of significance in shaping the current landscape of AI and NLP, enabling more sophisticated and accurate language-related tasks (LE et al., 2022; CHEN et al., 2023; DEVLIN et al., 2019).

2.4.4 GenomicBERT

BERT has been applied as a data driven approach to preprocess raw biological sequence data, and has proved as a viable tool for the job. Some authors have developed and published the tools to utilize a model called genomicBERT (CHEN et al., 2023), a model that have learned appropriate embedding for long sequences, tokenizing them while at the same time retaining the biological context.

With a reduced vocabulary size and handling cases of out-of-vocabulary, it was effective on it's proposal (CHEN et al., 2023), while demonstrating the plausibility of NLP tools applied to bioinformatics.

2.4.5 BERT-Promoter

A promoter, the portion of DNA that starts and regulates the transcription process, is very important to molecular biology. A BERT model, called BERT-Promoter, was developed specifically to figure out promoters in the DNA code. Not only BERT-Promoter was capable of finding promoter regions, it has differentiated them by their activities, strong or weak (LE et al., 2022), showcasing the relevance and opportunities to apply techniques from natural language processing to the biological data.

2.4.6 Related work

The closest work we found was the ProteinBERT (BRANDES et al., 2022), an model capable of predicting protein structure, biophysical attributes and post-translational modifications, with state-of-the-art performance.

3 MATERIALS AND METHODS

3.1 TOOLS

Given the importance, impact, easy of use and popularity of the programming language Python in the field of machine learning and the widely used libraries like Numpy and SciPy (RASCHKA, 2015), we choose python as the foundation to build upon and conduct all experiments.

3.2 DATASET SELECTION AND PREPROCESSING

The success of any machine learning endeavour hinges on the quality and relevance of the training data. For this study, we will compile a comprehensive dataset containing both enzyme and non-enzyme amino acid sequences. Each sequence will be represented in a standard string containing valid FASTA characters, that will be collected from The UniProt Consortium. An example of a FASTA file collected from UniProt Consortium is presented in annex A (CONSORTIUM, 2022).

The dataset will be split into amino acid sequences that codifies enzymes and non-enzymes given by the catalytic activity section from UnitProt, while being partitioned into training, validation, and test subsets as usual in machine learning procedure.

3.3 TRANSFORMER ENCODER ADAPTATION

The encoder segment of the Transformer architecture, as shown in Figure 1, was chosen for its innate ability to capture intricate dependencies within sequential data (VASWANI et al., 2023; DEVLIN et al., 2019; YANG et al., 2020). This segment consists of a stack of multi-head self-attention mechanisms and feedforward neural networks. In adapting the encoder to our task, we regarded each amino acid in the sequence as an analogous token in natural language processing.

The output of the encoder block will be passed through a linear activation layer and a softmax function to make the final binary classification, as can be seen in Figure 2. We thus discard the decoder, since our problem formulation does not incur the need for a recurring input based on previous output.

3.4 MODEL SIZE

The size of our model will be based both our computing budget and the amount of tokens we will be collecting, as following the elucidations emerged from the *Chinchilla* paper

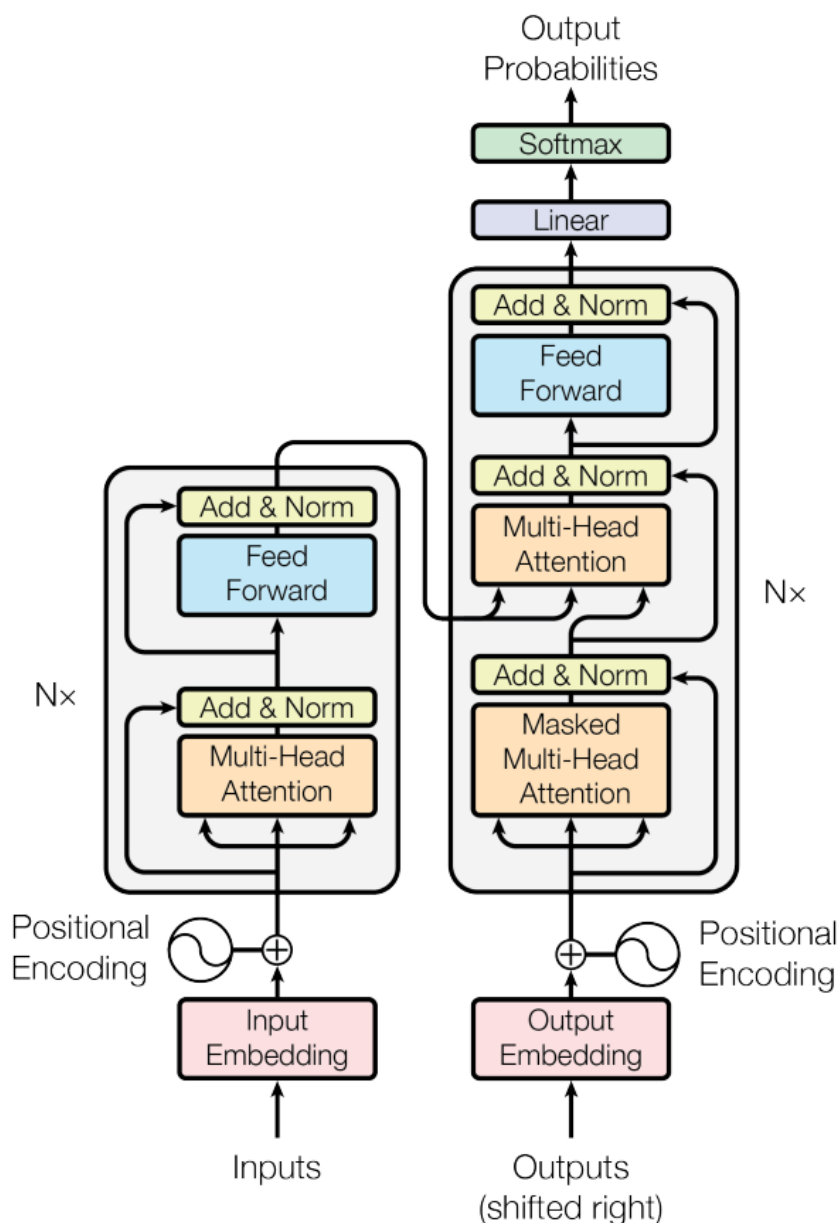


Figure 1 – The original Transformer architecture (VASWANI et al., 2023).

(HOFFMANN et al., 2022) to train a model optimally based on the amount of tokens at disposal and computing budget.

3.5 MODEL TRAINING

The amino acid sequences will be encoded using a byte-pair encoding with around 22 vocabulary token, to cover all aminoacids, padding and starting tokens. The input sequence tokens will be either padded or truncated according to the architecture input requirements.

The model performance will be optimised using cross-entropy loss, minimising the divergence between predicted and true labels. Training will follow the standard backpropagation and gradient descent. To mitigate overfitting, dropout regularisation will be employed during

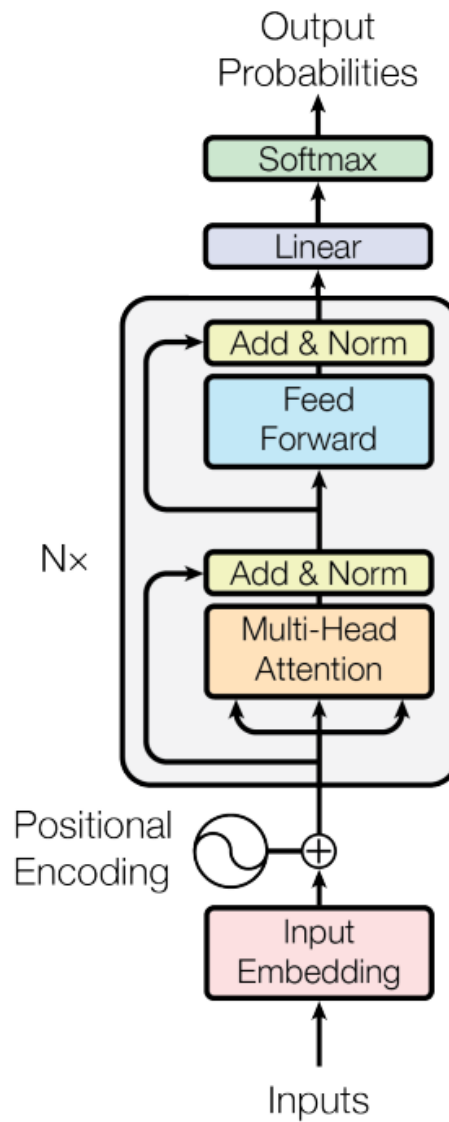


Figure 2 – Our architecture adapted from the original Transformer (VASWANI et al., 2023). Notice that we do not use the decoder part of the architecture, given that our problem formulation does not need to insert the outputs as input again.

training, as described in the transformer architecture for the base model (VASWANI et al., 2023) and used in other architectures (YANG et al., 2020).

3.6 EVALUATION AND VALIDATION

To assess the model efficacy, we will conduct cross-validation tests. Given that our model is tasked with only classification, it's performance will be evaluated using standard metrics including accuracy, precision, recall, and F1-score as a final metric for easily comparable results.

3.7 SCHEDULE

The schedule was divided based on the theoretical work and practical experiments. For the literature review, starting from 1st Oct. 2023 to 31st Mar 2024, the dissertation writing from 1st Nov 2023 to 31st May 2024.

The data collection starts from 1st Jan 2024 to 31st Mar 2024, the model development from 1st Feb 2024 to 31st Mar 2024, while the experiments from 15th Feb 2024 to 15th May 2024. The results analyses from 1st Apr 2024 to 31st May 2024, as seen in Figure 3.

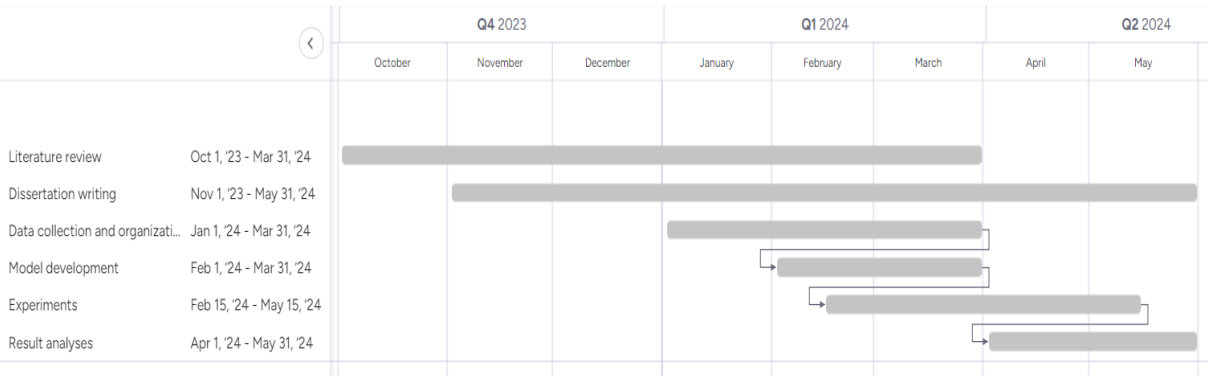


Figure 3 – The planned schedule. Made with monday.com (software as a service) work management web application.

4 EXPECTED RESULTS

We expect to organise a large dataset of aminoacids and classification for enzyme / not enzyme, to be used in future research projects, especially those in applying NLP techniques to bioinformatics data.

Our experimental results are expected to validate the robust capability of the Transformer encoder in classifying enzyme and non-enzyme sequences. The model is expected to achieve impressive accuracy, with precision and recall rates. The architecture's capacity to capture both local and global dependencies within amino acid sequences are expected to prove as an advancement in the automated sequence annotation classification performance field.

If our time budget and the size of the model requires, we plan to test other models and architectures like DistilBERT (SANH et al., 2020), for it's capability of compressing the model into a more resource usage friendly version of itself, expanding the use cases of such a model.

The model could be used as a filter, allowing future work to focus on the actual functional properties separately of both enzymes coding sequences, and non-enzyme sequences.

BIBLIOGRAPHY

- BARRAULT, Loïc et al. Seamless4t-massively multilingual & multimodal machine translation. **arXiv preprint arXiv:2308.11596**, 2023. pages 19
- BENSON, Dennis A et al. Genbank. **Nucleic acids research**, Oxford University Press, v. 41, n. D1, p. D36–D42, 2012. pages 15
- BRANDES, Nadav et al. ProteinBERT: a universal deep-learning model of protein sequence and function. **Bioinformatics**, v. 38, n. 8, p. 2102–2110, 02 2022. ISSN 1367-4803. Disponível em: <<https://doi.org/10.1093/bioinformatics/btac020>>. pages 9, 21
- CHEN, Tyrone et al. genomicbert and data-free deep-learning model evaluation. **bioRxiv**, Cold Spring Harbor Laboratory, p. 2023–05, 2023. pages 9, 21
- CHOI, Kristy et al. Encoding musical style with transformer autoencoders. In: III, Hal Daumé; SINGH, Aarti (Ed.). **Proceedings of the 37th International Conference on Machine Learning**. PMLR, 2020. (Proceedings of Machine Learning Research, v. 119), p. 1899–1908. Disponível em: <<https://proceedings.mlr.press/v119/choi20b.html>>. pages 19
- CONSORTIUM, The UniProt. UniProt: the Universal Protein Knowledgebase in 2023. **Nucleic Acids Research**, v. 51, n. D1, p. D523–D531, 11 2022. ISSN 0305-1048. Disponível em: <<https://doi.org/10.1093/nar/gkac1052>>. pages 10, 22
- COPELAND, Robert A. **Enzymes: a practical introduction to structure, mechanism, and data analysis**. [S.l.]: John Wiley & Sons, 2023. pages 9
- CUNNINGHAM, Fiona et al. Ensembl 2022. **Nucleic acids research**, Oxford University Press, v. 50, n. D1, p. D988–D995, 2022. pages 15
- DEVLIN, Jacob et al. **BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding**. 2019. pages 20, 21, 22
- DILL, Ken A.; MACCALLUM, Justin L. The protein-folding problem, 50 years on. **Science**, v. 338, n. 6110, p. 1042–1046, 2012. Disponível em: <<https://www.science.org/doi/abs/10.1126/science.1219021>>. pages 19
- DOSOVITSKIY, Alexey et al. **An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale**. 2021. pages 9, 18
- DRESZER, Timothy R et al. The ucsc genome browser database: extensions and updates 2011. **Nucleic acids research**, Oxford University Press, v. 40, n. D1, p. D918–D923, 2012. pages 15
- HOFFMANN, Jordan et al. **Training Compute-Optimal Large Language Models**. 2022. pages 23
- JONES, David T; THORNTON, Janet M. The impact of alphafold2 one year on. **Nature methods**, Nature Publishing Group US New York, v. 19, n. 1, p. 15–20, 2022. pages 20
- JUMPER, John et al. Highly accurate protein structure prediction with alphafold. **Nature**, Nature Publishing Group, v. 596, n. 7873, p. 583–589, 2021. pages 9, 15, 19

- KREUK, Felix et al. **AudioGen: Textually Guided Audio Generation**. 2023. pages 9, 18, 19
- LAKEW, Surafel M.; CETTOLO, Mauro; FEDERICO, Marcello. **A Comparison of Transformer and Recurrent Neural Networks on Multilingual Neural Machine Translation**. 2018. pages 9, 18, 19
- LE, Nguyen Quoc Khanh et al. Bert-promoter: An improved sequence-based predictor of dna promoter using bert pre-trained model and shap feature selection. **Computational Biology and Chemistry**, Elsevier, v. 99, p. 107732, 2022. pages 9, 21
- Authors' preface. In: PALMER, Trevor; BONNER, Philip L. (Ed.). **Enzymes (Second Edition)**. Second edition. Woodhead Publishing, 2011. p. xiv–xv. ISBN 978-1-904275-27-5. Disponível em: <<https://www.sciencedirect.com/science/article/pii/B978190427527500246>>. pages 9, 11, 12, 13, 14, 19
- RAFFEL, Colin et al. Exploring the limits of transfer learning with a unified text-to-text transformer. **CoRR**, abs/1910.10683, 2019. Disponível em: <<http://arxiv.org/abs/1910.10683>>. pages 20
- RASCHKA, Sebastian. **Python machine learning**. [S.l.]: Packt publishing ltd, 2015. pages 22
- ROBERTSON, Angus J. et al. Concordance of x-ray and alphafold2 models of sars-cov-2 main protease with residual dipolar couplings measured in solution. **Journal of the American Chemical Society**, v. 143, n. 46, p. 19306–19310, 2021. PMID: 34757725. Disponível em: <<https://doi.org/10.1021/jacs.1c10588>>. pages 20
- RUST, Alistair G.; MONGIN, Emmanuel; BIRNEY, Ewan. Genome annotation techniques: new approaches and challenges. **Drug Discovery Today**, v. 7, n. 11, p. S70–S76, 2002. ISSN 1359-6446. Disponível em: <<https://www.sciencedirect.com/science/article/pii/S1359644602022894>>. pages 9, 14, 16
- SANH, Victor et al. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. 2020. pages 26
- SCHUSTER, Mike; PALIWAL, Kuldip K. Bidirectional recurrent neural networks. **IEEE transactions on Signal Processing**, IEEE, v. 45, n. 11, p. 2673–2681, 1997. pages 17
- SHU, Chang et al. **SideRT: A Real-time Pure Transformer Architecture for Single Image Depth Estimation**. 2022. pages 9, 18
- SOH, Jung; GORDON, Paul MK; SENSEN, Christoph W. **Genome annotation**. [S.l.]: CRC Press, 2012. pages 9, 14, 15
- SONG, Yang et al. **Consistency Models**. 2023. pages 18
- TOUVRON, Hugo et al. Llama 2: Open foundation and fine-tuned chat models. **arXiv preprint arXiv:2307.09288**, 2023. pages 9, 18, 19
- TSOMPANA, Maria; BUCK, Michael J. Chromatin accessibility: a window into the genome. **Epigenetics & chromatin**, BioMed Central, v. 7, n. 1, p. 1–16, 2014. pages 15
- VASWANI, Ashish et al. **Attention Is All You Need**. 2023. pages 9, 10, 18, 22, 23, 24

WANG, Zilong; WAN, Zhaohong; WAN, Xiaojun. Transmodality: An end2end fusion method with transformer for multimodal sentiment analysis. In: **Proceedings of The Web Conference 2020**. New York, NY, USA: Association for Computing Machinery, 2020. (WWW '20), p. 2514–2520. ISBN 9781450370233. Disponível em: <<https://doi.org/10.1145/3366423.3380000>>. pages 9, 18

YANG, Zhilin et al. **XLNet: Generalized Autoregressive Pretraining for Language Understanding**. 2020. pages 22, 24

ZHIQIANG, Wang; JUN, Liu. A review of object detection based on convolutional neural network. In: IEEE. **2017 36th Chinese control conference (CCC)**. [S.l.], 2017. p. 11104–11109. pages 9, 16

Annex

ANNEX A – EXAMPLE OF A FASTA FILE

```

>sp|P09848|LPH_HUMAN Lactase/phlorizin hydrolase OS=Homo sapiens OX=9606
GN=LCT PE=1 SV=3
MELSWHVVFIALLSFSCWGS DWESDRNFISTAGPLTNDLLHNL SGLLDQSSNFVAGDKD
MYVCHQPLPTFLPEYFSSLHASQITHYKVFLSWAQLLPAGSTQNPDEKTVQCYRRLKAL
KTARLQPMVILHHQTLPASTLRRTEAFADLFADYATFAFHSFGDLVGIWFTFSDLEEVIK
ELPHQESRASQLQTLSDAHRKAYEIYHESYAFQGGKLSVVLRAEDIPELLLEPPISALAQ
DTVDFLSLDLSYECQNEASLRQKLSKLQTIIEPKVKVFIFNLKLPDCPSTMKNPASLLFSL
FEINKDQVLTIGFDINEFLSCSSSSKSMSCSLTGSLALQPDQQQDHETDSSPASAYQ
RIWEAFANQSRAERDAFLQDTFPEGFLWGASTGAFNVEGGWAEGGRGVSIWDPRRPLNTT
EGQATLEVASDSYHKVASDVALLCGLRAQVYKFSISWSRIFPMGHGSSPSLPGVAYYNKL
IDRLQDAGIEPMATLFHWDLPQALQDHGGWQNESVVDAFLDYAAFCFSTFGDRVKLWVTF
HEPWVMSYAGYGTGQHPPGISDPGVASFVAHLVLKAHARTWHHYNSHHRPQQQGHVGIV
LNSDWAEP LSPERPEDLRASERFLHFMLGWFAHPVFVDGDYPATLRTQIQQMNRQC SHPV
AQLPEFTEAEKQLLKGSADFLGLSHYTSRLISNAPQNTCIPSYDTIGGFSQHVNHVWPQT
SSSWIRVVPWGIRRLQFVSLEYTRGKVPIYLAGNGMPIGESENLFDDSLRVDYFNQYIN
EVLKAIKEDSVDVRSYIARSLIDGFEGPSGYSQRFGLHHVNFSDSSKSRTPRKSAYFFTS
IIEKNGFLT KGAKRLLPPNTVNLPSKVRAFTFPSEVPSKAKVVWEKFSSQPKFERDLFYH
GTFRDDLFWGVSSSAYQIEGAWDADGKGPSIWDNFTHTPGSNVKDNATGDIACDSYHQLD
ADLNMLRALKVKAYRFSISWSRIFPTGRNSSINSHGV DYYNRLINGLVASNIFPMVTLFH
WDL PQALQDIGGWENPALIDLFD SYADFCFQTFGDRV KFWMTFN EPMYLA WLGYGSGEFP
PGVKDPGWAPYRIAHAVIKAHARVYHTYDEKYRQE QKGVISLSLSTHWAEPKSPGVPRDV
EAADRMLQFSLGWFAHPIFRNGDYPDTMKWKVGNRSELQHLATSRLPSFTEEEKR FIRAT
ADVFC LNTYYSRIVQHKT PRLNPPSYEDDQEMAEEDPSWPSTAMNRAAPWGTRRLLNWI
KEEYGDIPYITENG VGLTNPNTEDTD RIFYHKTYINEAL KAYRLDGIDLRGYVAWSLMD
NFEWLN GYTVKFGLYHVD FNNTNRPR TARASARYYTEVITNNGMPLAREDEF LYGRFPEG
FIWSAASAAYQIEGAWRADGKGLSIWDTFSHTPLRVENDAIGDVACDSYHKIAEDLVTLQ
NLGVSHYRFSISWSRILPDGTTRYINEAGLNYVRLIDTLLAASIQQVTIYHWDLPQTL
QDVGGWENETIVQRFKEYADVLFQRLGDKVKFWITLNEPFVIAYQGYGYGTAAPGVSNRP
GTAPYIVGHNLIKAHAEAWHLYNDVYRASQGGVISITISSDWAEP RDP SNQEDVEAARRY
VQFMGGWFAHPIFKNGDYNEVMKTRIRD RSLAAGLNKSRLPEFTESEKRRINGTYDFFGF
NHYYTTVLAYNLNYATAISSFDADRGVASIADRSWPD SGFWLKMTPFGFRILNLWLKEEY
NDPPIYVTENGVSQREETDLNDTARIYYLR TYINEALKAVQDKVDLRGYTVWSAMD NFEW
ATGFSERFGLHFVNYS DPSLPRIPKASAKFYASVVR CNGFDPATGPHACLHQP DAGPTI
SPVRQEEVQFLGLMLGTTEAQTALYVLFSLVLLGVCGLAFLSYKYCKRSKQ GKTQRSQQE
LSPVSSF

```