الجمهورية الجزائرية الديمقراطية الشعبية
People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research

Constantine 1 Frères Mentouri University
Faculty of Natural and Life Sciences

جامعة قسنطينة 1 الإخوة منتوري
كلية علوم الطبيعة والحياة

Department of Applied Biology

قسم البيولوجيا التطبيقية

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master

Domain : Natural and Life Sciences

Field : Biotechnology

Specialization : Bioinformatics

N° d'ordre :
N° de série :

Title :

Prediction of antimicrobial peptide function based on a fine-tuned large language model

Submitted by:
OUFEROUKH Oussema
LAZOUNE Dalal

Sustained on: 10/06/2024

Board of Examiners :

Chairperson : Dr. H. BOUHALOUF (MCB - Constantine 1 Frère Mentouri University).

Supervisor : Dr. H. CHEHILI (MCA - Constantine 1 Frère Mentouri University).

Examiner : Dr. Y. MEZIANI (MAB - Constantine 1 Frère Mentouri University).

Academic year
2023 - 2024

Acknowledgements

First and foremost, we sincerely thank God for blessing us with the perseverance, strength, and patience needed to bring this work to fruition.

We extend our heartfelt gratitude to our supervisor, Dr. CHEHILI Hamza, for his unwavering guidance, patience, and encouragement. His expertise and insights have been invaluable, and his belief in our potential has been a constant source of motivation. We are deeply grateful for his presence, availability, and involvement in helping us complete this work.

Our thanks also go to the members of the jury, Dr. H BOUHALOUF and Dr. Y MEZIANI, for their interest in our research, agreeing to evaluate our thesis, and enriching it with their contributions. We would also like to thank all our teachers who have provided our training throughout these years. Their efforts have been immensely beneficial in carrying out this work

To our family—our parents, brothers, and sisters—thank you for your endless love and support.

Our parents, your faith in us has been a beacon of strength, guiding us through the toughest times. To our siblings, thank you for being our cheerleaders and grounding force.

A special thanks to our friends and colleagues who have provided not only academic support but also emotional sustenance. Your camaraderie has made this journey bearable and often joyful. Thank you for all the moments we shared.

This thesis is not just the culmination of our academic efforts, but a testament to the collective support, love, and encouragement from all of you. We are eternally grateful.

Thank you.

Dalal *Oussema*

Abstract

Antimicrobial peptides (AMPs) play a crucial role in the innate immune system by effectively combating disease-causing pathogens. The rapid increase of drug-resistant infections poses serious challenges to current antimicrobial therapies. Traditional wetlab experimentation for AMP identification is known to be cost-intensive. Therefore, the incorporation of efficient computational tools becomes imperative in the preemptive identification of optimal AMP candidates prior to in vitro experimentation.

In this study, we introduce AMP-Gemma, a fine-tuned large language model tailored for the precise prediction of antimicrobial peptides. Through our model, AMP-Gemma, we achieved an exceptional accuracy rate of 86.24% in accurately predicting peptides with antimicrobial activity, surpassing the performance of existing machine learning and deep learning methodologies for AMP prediction.

Furthermore, we have developed a user-friendly chatbot specializing in AMPs, designed to cater to the needs of the biological community. This innovative tool aims to facilitate access to information and streamline communication within the field of antimicrobial peptide research.

Keywords :

Antimicrobial peptides (AMPs), AMP-Gemma, Large language models (LLMs), Machine learning, Deep learning, Antimicrobial activity.

Résumé

Les peptides antimicrobiens (AMPs) jouent un rôle crucial dans le système immunitaire inné en combattant efficacement les agents pathogènes responsables des maladies. L'augmentation rapide des infections résistantes aux médicaments pose des défis sérieux aux thérapies antimicrobiennes actuelles. L'expérimentation traditionnelle en laboratoire humide pour l'identification des AMPs est connue pour être coûteuse. Par conséquent, l'intégration d'outils informatiques efficaces devient impérative pour l'identification préventive des candidats AMPs optimaux avant l'expérimentation in vitro.

Dans cette étude, nous présentons AMP-Gemma, un modèle de grande langue finement ajusté pour la prédiction précise des peptides antimicrobiens. Grâce à notre modèle, AMP-Gemma, nous avons atteint un taux de précision exceptionnel de 86.24% dans la prédiction précise des peptides ayant une activité antimicrobienne, surpassant les performances des méthodologies existantes d'apprentissage automatique et d'apprentissage profond pour la prédiction des AMPs.

De plus, nous avons développé un chatbot convivial spécialisé dans les AMPs, conçu pour répondre aux besoins de la communauté biologique. Cet outil innovant vise à faciliter l'accès à l'information et à simplifier la communication dans le domaine de la recherche sur les peptides antimicrobiens.

Mots clés :

Peptides antimicrobiens (AMPs), AMP-Gemma, Modèles de grande langue (LLMs), Apprentissage automatique, Apprentissage profond, Chatbot, Bactéries résistantes aux antibiotiques, Activité antimicrobienne.

ملخص

تؤدي الببتيدات المضادة للميكروبات (AMPs) دوراً حيوياً في الجهاز المناعي من خلال مكافحة مسببات الأمراض بفعالية. يشكل الارتفاع السريع في الإصابات المقاومة للأدوية تحديات خطيرة للعلاجات المضادة للميكروبات الحالية. ومن المعروف أن التجارب التقليدية في المختبرات الرطبة لتحديد الببتيدات المضادة للميكروبات AMPs مكلفة للغاية. لذلك، يصبح دمج الأدوات الحاسوبية الفعالة أمراً حتمياً في التعرف الاستباقي على أفضل مرشحي هذه الببتيدات قبل التجارب المختبرية.

في هذه الدراسة، نقدم AMP-Gemma، وهو نموذج لغوي كبير مُحسَّن بدقة للتنبؤ الببتيدات المضادة للميكروبات. من خلال نموذجنا AMP-Gemma، حققنا معدل دقة استثنائي بلغ 86.24٪ في التنبؤ الدقيق بالببتيدات ذات النشاط المضاد للميكروبات، متفوقاً على أداء منهجيات التعلم الآلي والتعلم العميق الحالية في التنبؤ الببتيدات المضادة للميكروبات .

علاوة على ذلك، قمنا بتطوير روبوت محادثة سهل الاستخدام متخصص في الببتيدات المضادة للميكروبات، مصمم لتلبية احتياجات المجتمع البيولوجي. يهدف هذا الأداة المبتكرة إلى تسهيل الوصول إلى المعلومات وتبسيط التواصل في مجال أبحاث الببتيدات المضادة للميكروبات.

الكلمات المفتاحية:

الببتيدات المضادة للميكروبات (AMPs)، AMP-Gemma، النماذج اللغوية الكبيرة (LLMs)، التعلم الآلي، التعلم العميق، النشاط المضاد للميكروبات.

List of Tables

List of Figures

ACRONYMS

AMPs: Antimicrobial peptides.

AMR: Antimicrobial resistance.

ACPs: Anticancer peptides.

NLP: Natural Language Processing.

AI: Artificial Intelligence.

ML: Machine Learning.

DL: Deep Learning.

LLMs: Large Language Models.

APD3: Antimicrobial Peptide Database.

MIC: Minimal Inhibitory Concentration.

MBC: Minimal Bactericidal Concentrations.

FDA: Food and Drug Administration.

ANN: Artificial Neural Networks.

RF: Random Forest.

SVM: Support Vector Machine.

AAC: Amino Acid Composition.

PseAAC: Pseudo Amino Acid Composition.

QM: Quantitative Matrix.

GPUs: Graphics Processing Unit.

TPUs: Tensor Processing Units.

MLM: Masked Language Modeling.

RoPE: Rotary Positional Embeddings.

UniProt: The Universal Protein Resource**.**

DRAMP: Data Repository of Antimicrobial Peptides.

HPC: High Performance Computing.

PEFT: Parameter-Efficient Fine-Tuning.

CSV: Comma-Separated Values.

ACC: Accuracy.

SN: Sensetivity or Recall.

F1: F1-Score.

TP: True Positive.

TN: True Negative.

FP: False Positive.

FN: False Negative.

XGBoost (XGB): Extreme Gradient Boosting.

MNB: Multinomial Naive Bayes.

LR: Logistic Regression.

KNN: KNearest Neighbor.

MLP: MultiLayer Perceptron (MLP).

ESM: Evolutionary Scale Model.

GPT-3.5 : Generative Pre-trained Transformer.

Contents

Chapitre II : Materials and Methods

Chapitre III : Results and discussion

INTRODUCTION

Introduction

In recent years, antimicrobial resistance has emerged as a significant global threat. The World Health Organization (WHO) has extensively announced the alarming global increase in resistance to conventional antimicrobials as a potential and serious risk to public health [1]. Antimicrobial resistance (AMR) is recognized as one of the major global threats to public health. In 2019, bacterial AMR directly caused 1.27 million global deaths and contributed to an additional 4.95 million deaths [2].

Hence, antibiotic resistance poses a serious challenge and has necessitated the development of new alternative molecules less susceptible to bacterial resistance. Antimicrobial peptides (AMPs) have garnered significant attention as potential next-generation antibiotics. They are bioactive small proteins that are naturally produced by all living organisms, serving as the primary defense against various human infections such as fungi, viruses, and bacteria [3]. Consequently the discovery of peptides with antimicrobial properties is essential for the advancement of novel therapeutics [4].

However, the indiscriminate and prolonged use of antibiotics, especially in developing countries, in both human and veterinary medicine, as well as in agriculture, has contributed to the development and spread of drug-resistant microorganisms [5], and the traditional method of discovering these bioactive peptides is a lengthy and laborious process. Streamlining this process is essential for the development of new therapeutics with significant health benefits.

On the other hand, bioinformatics, the field that integrates rapid computational biology, has emerged as the anticipated solution to this issue. Bioinformatics has been introduced as the means to efficiently search for bioactive peptides [6].

Over the past few decades, Machine Learning (ML) and Deep Learning models have gained popularity in a wide range of real-world applications. One such application is the prediction of Antimicrobial peptides using AMPs datasets, and these models have demonstrated significant accuracy in predicting AMPs.

In 2023, there was a notable increase in the exploration of a different type of Machine Learning known as Large Language Models (LLMs) across different fields. Throughout the year, we examined the utilization of LLMs in diverse areas of bioinformatics and biomedical

informatics, including omics, genetics, biomedical text mining, drug discovery, biomedical image analysis, bioinformatics programming, and bioinformatics education.

The significant progress in generative models, particularly Large Language Models, has marked a revolutionary period in Natural Language Processing (NLP) [7]. LLMs exhibit exceptional capability in understanding and generating human-like text, playing a crucial role in various NLP tasks such as machine translation [8], commonsense reasoning [9], and coding assignments [10].

The notable advancements in language models, largely credited to transformers [11], enhanced computational power, and the abundance of extensive training data, have led to a groundbreaking shift. These progressions have facilitated the development of Large Language Models (LLMs) that can approximate human-level performance across a range of tasks [12].

The recent advancements in language models have provided the protein modeling community with a powerful tool that utilizes transformers to depict protein sequences as text. This innovation allows for the prediction of sequence-to-property for peptides without the need for explicit structural data [13].

In this work, we present AMP-GEMMA, a language model that predicts the Antimicrobial peptide properties using only amino acid sequences as the input. Our model is the result of fine-tuning the open lightweight language model Gemma.

The aim is to initiate new approaches for identifying Antimicrobial peptides and in a short period of time and with LLMs and improve accuracy which make the discovery and developing novel therapeutics more specific and effective.

As a second contribution, we introduce a chat model based on Gemma for generating specific answers about antimicrobial peptides. This chat model provides researchers with a comprehensive analysis of the peptides, including their structure, functional properties, and classification. This facilitates and advances the scientific research process on AMPs through the use of an advanced and precise tool, which is our chat model.

This thesis is divided into three chapters :

- The first chapter delves into Antimicrobial peptides (threats, origins, classification of AMPs), The AMPs as in addressing microbial infections induced by different factors, and explores Artificial Intelligence (AI) concepts including its subfields and components such as Machine Learning (ML), Deep Learning (DL), and Large Language Models (LLMs).
- Chapter 2 outlines the experimental work conducted in the project.

- In Chapter 3, the main findings of our research are described, while the discussion section interprets the results to highlight the significance of the findings.

CHAPTER I :

Bibliographic Synthesis

1. Antimicrobial Peptides

Antimicrobial peptides (AMPs) are oligopeptides characterized by a variable number of amino acids, typically ranging from five to over a hundred, and known for their antimicrobial properties [14]. They exhibit diverse functions such as antibacterial, antifungal, antiviral, and in some cases, even anticancer activities [7], and they represent an expanding group of natural and synthetic peptides that target a broad range of organisms, including viruses, bacteria, fungi, and parasites [14]. This has led to a search for alternative methods of treating microbial infections.

The field of antimicrobial peptides is expanding for several key reasons: (1) The critical requirement for new antimicrobial agents to combat drug-resistant pathogens, including superbugs, viruses, fungi, and parasites. (2) The quest to enhance our comprehension of the biological roles of natural AMPs in innate immunity. (3) The increasing fascination with microbiota shaped by AMPs [15].

Hence, it is crucial to discover potential AMP candidates and their characteristics to defend against resistant microorganisms .

1.1. Discovery of AMPs

The discovery of AMPs can be traced back to 1939, when Dubos extracted an antimicrobial agent from a soil Bacillus strain. This extract was shown to protect mice from pneumococcal infection. The following year, Hotchkiss and Dubos fractionated this extract and identified an AMP called gramicidin. Despite reported toxicity associated with intraperitoneal application, gramicidin was found to be effective for topical treatment of wounds and ulcers. In 1941, another AMP called tyrocidine was discovered and found to be effective against both Gram-negative and Gram-positive bacteria. However, tyrocidine exhibited toxicity to human blood cells. In the same year, another AMP was isolated from a plant called Triticumaestivum, which was later named purothionin and found to be effective against fungi and some pathogenic bacteria [14].

The discovery of antimicrobial peptides in later years has continued until today. Some of the important antimicrobial peptides discovered to date include Plant Kalata B1 in 1973, cecropins produced by insects in 1980, magainin from frogs in 1987, and defensins produced by human and mammalian cells in 1985 [16].

1.2. The source of AMPs

Natural antimicrobial peptides (AMPs) are present in both prokaryotes, such as bacteria, and eukaryotes [14]. The sources of AMPs can be divided into mammals (human host defense peptides account for a large proportion), amphibians, microorganisms, and insects according to statistical data in APD3, AMPs found in oceans have also attracted widespread attention [17].

1.3. Classification of AMPs

There are various classifications of antimicrobial peptides based on: The biosynthetic machines, biological sources, biological functions, molecular properties, covalent bonding patterns, 3 dimensional structures, molecular targets [16]. The classification of AMPs based on their target, and mode of action, for natural AMPs, from eukaryotes, especially mammals as the follow :

1.3.1. Antiviral Peptides

Antiviral AMPs neutralize viruses by incorporating into either the viral envelope or the host cell membrane [14]. Previous studies have demonstrated that antiviral AMPs can target both enveloped RNA and DNA viruses [18,19], and the AMPs can integrate into viral envelopes, causing membrane instability and rendering the viruses incapable of infecting host cells [20,21].

1.3.2. Antibacterial Peptides

Antibacterial AMPs are the most studied AMPs to date and most of them are cationic AMPs, which target bacterial cell membranes and disrupt the lipid bilayer structure [22,23]. The majority of these AMPs are also amphipathic, possessing both hydrophilic and hydrophobic domains [14]. These structural features enable AMPs to bind to lipid components via their hydrophobic regions and to phospholipid groups via their hydrophilic regions [24].

1.3.3. Antifungal Peptides

Similar to antibacterial AMPs, antifungal peptides can kill fungi by targeting either the cell wall [25, 26] or intracellular components [27]. However, the bacterial membrane and fungal cell wall have different compositions, and this binding ability helps AMPs efficiently target fungal cells [14]. Antifungal AMPs that target the cell wall kill fungal cells by

disrupting membrane integrity [28,29], increasing plasma membrane permeability [30], or forming pores directly [31].

1.3.4. Antiparasitic Peptides

Antiparasitic peptides constitute a smaller group compared to the other three AMP classes [14]. The first reported antiparasitic peptide is magainin, which can kill Paramecium caudatum [32]. Despite some parasitic microorganisms being multicellular, the mode of action of antiparasitic peptides is similar to that of other AMPs [14]. They kill cells by directly interacting with the cell membrane [32].

1.3.5. Anticancer Peptides

Several antimicrobial peptides have been studied as potential candidates for anticancer drugs [17], offering a new direction for cancer treatment [33]. These anticancer peptides (ACPs) are considered safer than synthetic drugs [34,35]. The mechanisms by which ACPs exhibit anticancer activity include (1) recruiting immune cells, such as dendritic cells, to kill tumor cells, (2) inducing necrosis or apoptosis in cancer cells, (3) inhibiting angiogenesis to starve tumors and prevent metastasis, and (4) activating regulatory functional proteins to interfere with the gene transcription and translation in tumor cells [36].

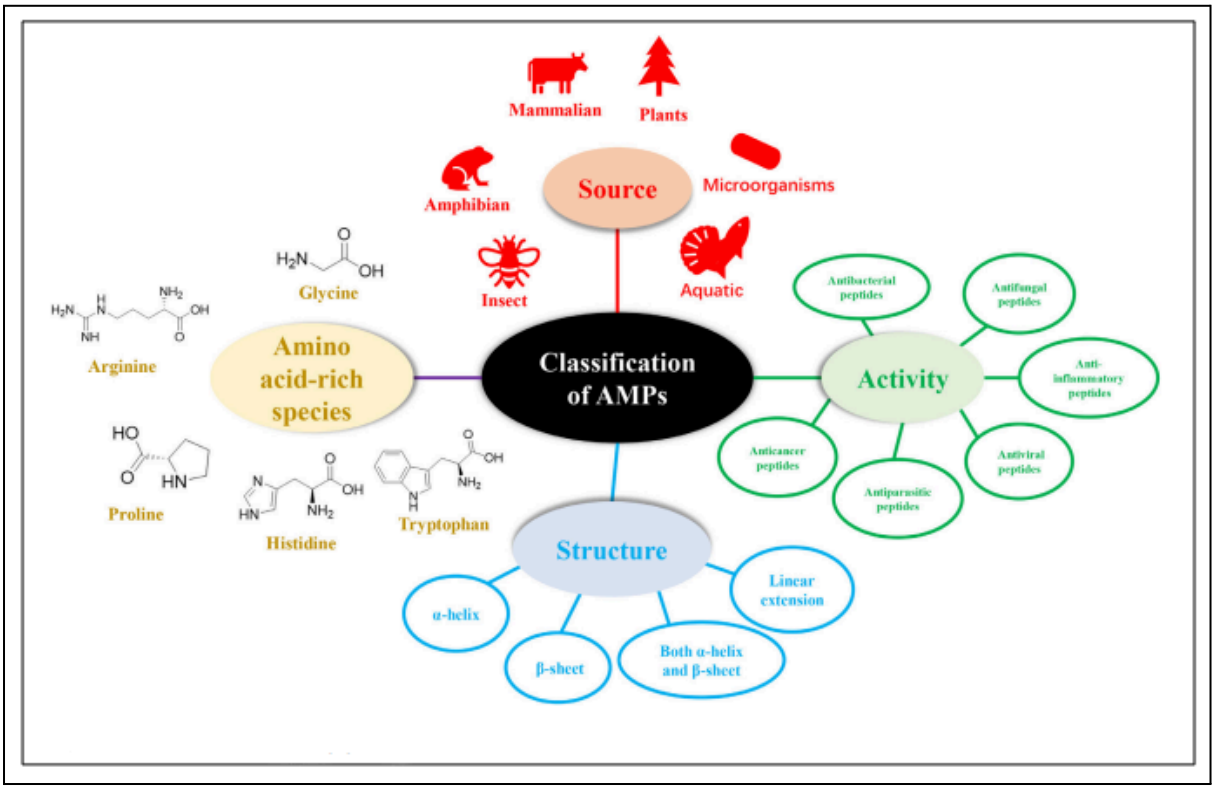

Figure 01 : Antimicrobial Peptides Classification [18]

1.4. Current progress and application of antimicrobial peptides

Some antimicrobial peptides are used as therapeutic agents , they have desirable effect in reducing the minimal inhibitory concentration (MIC) and minimal bactericidal concentrations (MBC) of those antibiotics because of their synergic activity in association with other antibiotics [17] .

The application of antimicrobial peptides (AMPs) as anticancer peptides (ACPs) in cancer therapy, either alone or in combination with traditional drugs, is being recognized as a promising therapeutic approach worthy of investigation [37]. Several applications have been demonstrated for therapeutic use, including: as standalone anti-infective agents, in combination with conventional antibiotics or antivirals to enhance additive or synergistic effects, as immune stimulatory agents to boost natural innate immunity, and as endotoxin neutralizing agents to prevent potentially fatal complications associated with bacterial virulence factors that cause septic shock [38].

In medicine, antimicrobial peptides play a crucial role in regulating pro-inflammatory reactions, recruiting cells, stimulating cell proliferation, promoting wound healing, modifying gene expression, and combating cancer cells, thereby contributing to immune regulation in human skin, respiratory infections, and inflammatory diseases [39]. The application of AMPs

in various medical fields, such as dental care, surgical infection prevention, wound healing, and ophthalmology, is currently evolving. However, only three AMPs, namely gramicidin, daptomycin, and colistin, have been approved by the FDA [18].

In the realm of food, AMPs exhibit potent inhibitory effects against common bacteria and fungi found in food. Many AMPs also demonstrate resistance to acids, alkalis, and high temperatures, although they can be readily hydrolyzed by proteases in the human body [18]. Examples include Enterocin AS-48, used for preserving cider, fruit and vegetable juices, and Enterocin CCM4231, employed for preserving soy milk [40, 41].

Peptide	**Description**	**Phase of clinical trial**	**Companies**	**Reference**
Magainin	a 22-amino-acid linear antimicrobial peptide, isolated from the skin of the African clawed frog (Xenopuslaevis)	phase III	Dipexium Pharma (White Plains, New York)/MacroChem/ Genaera for Diabetic foot ulcers	[42]
Omiganan	a synthetic cationic peptide derived from indolicidin	phase II	BioWest Therapeutics/Maru ho (Vancouver) for Rosacea	[42]
NVB302	a Class B lantibiotic	phase I	Novacta (Welwyn Garden City, UK) for C. difficile infection	[42]
Avidocin and purocin	a modified R-type bacteriocins from Pseudomonas aeruginosa	preclinical	Avid Biotics (S. San Francisco, California) and are narrow spectrum antibiotics for human health and food safety	[42]
IMX924	a synthetic 5-amino-acid peptide innate defense regulator, and and is effective in Gram-negative and Gram-positive bacterial infections and improves survival and reduces tissue damage	preclinical	Iminex (Coquitlam, British Columbia, Canada)	[42]

PAC-113, P-113	Histatin 5 derivative (12 amino acids) for Oral candidiasis	Phase IIb	General Biologicals Corporation	[18]
D2A21	Synthetic peptide for Burn wound infections	Phase III	Demegen	[18]
PXL01	Lactoferrin analog for Postsurgical adhesions	Phase III	ProMore Pharma	[18]

Table 01: Some of AMPs that are used in drug development.

1.5. Antimicrobial Peptides action mechanism

AMPs kill cells through various mechanisms, such as disrupting membrane integrity via interaction with negatively charged cell membranes, inhibiting proteins, DNA, and RNA synthesis, or interacting with specific intracellular targets [14].

Typically, an AMP is effective against only one class of microorganisms (e.g., bacteria or fungi) [43]. However, exceptions exist, and some AMPs are known to exhibit different modes of action against various types of microorganisms. For instance, indolicidin demonstrates the ability to combat bacteria, fungi, and HIV [14,44,45].

2. Artificial Intelligence (AI)

Now a computer system can perform tasks that typically rely on human intelligence, such as visual perception, speech recognition, decision-making, and language translation, thanks to Artificial Intelligence [46].

Artificial Intelligence (AI) can be defined as the study of mental and psychological capabilities through the utilization of different computational patterns and sequences [47].

It can also be described as the scientific endeavor to develop advanced machines, devices, and computer programs that analyze human intelligence [48] to solve the practical challenges that the world confronts us with.

The surge in AI can be linked to the combination of virtually limitless storage capacity and an abundance of data from various sources (such as images, text, transactions, and mapping data) a phenomenon known as the Big Data movement [46].

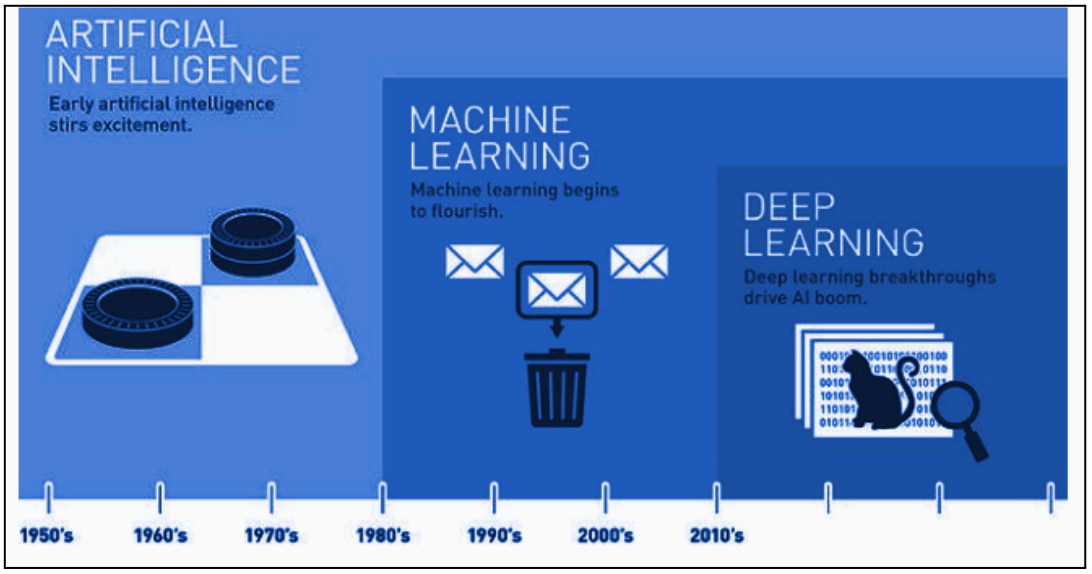

Figure 02 : Artificial Intelligence's Subsets [46]

3. Machine Learning (ML)

Machine learning, a subset of Artificial Intelligence (AI), equips computers with the capability to learn without explicit programming [47]. By enabling computers to intelligently perform specific tasks, machine learning systems can execute complex processes by learning from data, rather than adhering to pre-programmed rules [46].

Machine learning facilitates intelligent systems that can learn a specific function when provided with a particular dataset to learn from . In ML, tasks are typically categorized into broad groups based on how learning is acquired or how feedback on the learning process is provided to the developed system [46]. The machine learning algorithms are often categorized as being : Supervised, Un-supervised, Semi-supervised Learning, and Reinforcement learning [47].

4. Deep Learning (DL)

Deep learning is a subset of machine learning. In the realm of fluid mechanics, data-driven models are commonly constructed using Artificial Neural Networks (ANN) [47]. These ANNs draw inspiration from our knowledge of the biological structure of the brain and the intricate interconnections between neurons [49].

5. Large Language Models (LLMs)

Large Language Models (LLMs) is one of machine learning types, specifically engineered to process and produce natural language. Trained on extensive text data like books, articles, and websites, they have the capability to generate new text that is similar to the content of the training data [50].

Trained through unsupervised learning, LLMs learn directly from the data without requiring explicit labels or annotations. This enables them to grasp various patterns and structures within the training data and can be fine-tuned for particular tasks by providing them with a smaller set of task-specific data [50].

LLMs are based on transformers, a type of neural network architecture invented by Google. The transformer architecture is so powerful because of its ability to scale effectively, allowing us to train these models on massive text datasets. Transformers consist on encoder and decoder, the encoder encodes the input sequence and passes it to the decoder which learns how to decode the representations for relevant tasks.

The latest advancements in language models have provided the protein modeling field with a potent tool utilizing transformers to present protein sequences as text. This innovation allows for sequence-to-property prediction for peptides without the need for explicit structural data [13].

The advent of transformers and large language models (LLMs) has led to the development of novel deep learning architectures for modeling protein sequences, as amino acid sequences can be viewed as analogous to words and sentences in language. Particularly, the attention mechanism of LLMs enables them to capture both immediate and intricate connections among elements in diverse types of textual data. Consequently, this has sparked a rejuvenation in the field of bioinformatics, as protein sequences, similar to languages, demonstrate intricate interactions among amino acids [13].

By utilizing LLM and transformers, we can now harness sophisticated language modeling techniques to explore the roles of amino acids in the features of proteins [51].

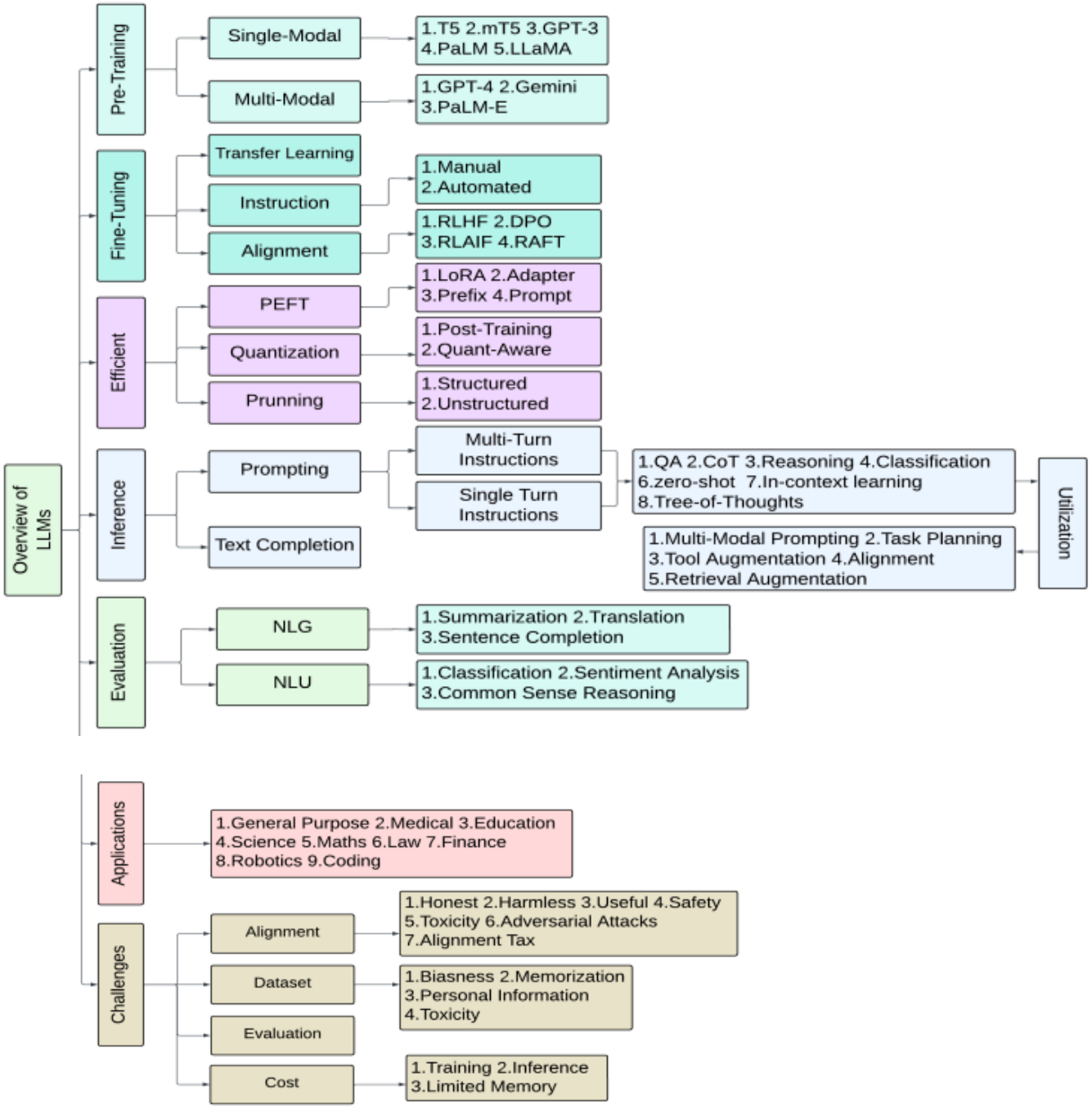

Figure 03 : A broader overview of LLMs, dividing LLMs into seven branches: 1. Pre-Training 2. Fine-Tuning 3. Efficient 4. Inference 5. Evaluation 6. Applications 7. Challenges [12]

6. Related works for predicting AMPs using artificial intelligence

The use of artificial intelligence and machine learning is increasingly prevalent. Recent developments in Artificial Intelligence (AI) are set to play a very essential role in the bioinformatics domain. Therefore, several AMP-related bioinformatics research that integrates diverse biological and chemical data with computational methods have been developed to identify candidate peptides with antimicrobial functions.

6.1. Machine learning methods

Various physicochemical properties, including amino acid composition (AAC), pseudo amino acid composition (PseAAC), charge, isoelectric point, hydrophobicity, polarity, and secondary structure, are used as input features in machine learning methods for predicting AMPs [52]. Algorithms such as random forest (RF) or support vector machine (SVM) are used to process these features. Machine learning methods previously proposed for AMP classification include :

6.1.1. CAMPr3

CAMPr3, an update to the existing CAMP database, is a repository containing sequences, structures, and family-specific signatures of prokaryotic and eukaryotic AMPs. It encompasses 10,247 sequences, 757 structures, and 114 family-specific signatures of AMPs. CAMPr3 offers detailed information on AMPs and their respective families, including sequences, structures, activity, signatures, as well as the source and target organisms [53].

6.1.2. iAMPpred

iAMPpred is an online prediction for AMP with improved accuracy by integrating compositional, physicochemical, and structural features into Chou's general PseAAC [54].

6.1.3. AmPEP

AmPEP , Sequence-based prediction of AMPs , utilizing distribution patterns of amino acid properties and the random forest algorithm [55] .

6.1.4. AntiBP2

AntiBP2 is an improved version of antibacterial peptide prediction tool that employs SVM, QM (quantitative matrix), and artificial neural network (ANN) algorithms [56].

6.1.5. CS-AMPPred

CS-AMPPred, Cysteine-Stabilized Antimicrobial Peptides Predictor, the revised version of the Support Vector Machine (SVM) model for predicting antimicrobial activity in cysteine-stabilized peptides relies on five sequence descriptors: indexes of alpha-helix and loop formation, as well as averages of net charge, hydrophobicity, and flexibility [57].

6.2. Deep learning-based methods

In recent years, numerous deep learning-based approaches have been created to categorize peptides as potential antimicrobial peptides (AMPs), including:

6.2.1. AMPscanner

AMPscanner is an advanced deep learning tool that enhances antimicrobial peptide recognition through the utilization of convolutional and recurrent layers that exploit primary sequence composition [58].

6.2.2. APIN

APIN, a method for antimicrobial peptide identification, employs a multi-scale convolutional network that surpasses the performance of existing models on two AMP datasets and the Antimicrobial Peptide Database (APD)3 benchmark dataset. It also outperforms the current state-of-the-art model in accuracy on an anti-inflammatory peptides (AIPs) dataset [59].

6.2.3. ACEP

ACEP is an innovative approach for AMPs prediction by automatically fusing features and embedding amino acids. This deep learning model is capable of acquiring amino acid embedding patterns, extracting sequence features automatically, and integrating heterogeneous information sources [60].

6.3. Large Language Models used in Bioinformatics tasks

Some LLMs have been created in the field of bioinformatics such as :

6.3.1. PeptideBERT

PeptideBERT is a language model based on Transformers designed for predicting peptide properties. This protein language model is customized for predicting key peptide characteristics such as hemolysis, solubility, and nonfouling. It utilizes the ProtBERT pretrained transformer model with 12 attention heads and 12 hidden layers, and it predicts peptide properties using only as input an amino acid sequences [13].

6.3.2. ProtTrans

ProtTrans is a Protein Language Model (pLM) that adapts the principles of Language Models from NLP by representing amino acids from protein sequences as tokens (equivalent to words in NLP) and treating complete proteins as sentences in Language Models. It provides state of the art pre-trained models for proteins and was trained on numerous GPUs from Summit and hundreds of Google TPUs utilizing diverse Transformer models [61].

6.3.3. Prot-BERT

Prot-BERT is a pretrained model for protein sequences that employs a masked language modeling (MLM) objective. Initially presented in [61] and subsequently released in [62], this model operates exclusively with uppercase amino acids, recognizing only amino acids represented by capital letters. Prot-BERT based on the Bert model, and is pretrained on an extensive collection of protein sequences in a self-supervised fashion [63].

CHAPTER II :

Materials and Methods

1. Materials

1.1. Large language model Gemma

In this work we have used Gemma, Google's latest series of four Large Language Models (LLMs) developed as part of the Gemini initiative. These models are available in two sizes: 2B and 7B parameters, each offering a base (pretrained) and an instruction-tuned variant. Both models use a head dimension of 256, and both variants utilize Rotary Positional Embeddings (RoPE) [64].

Designed to function efficiently on a range of consumer hardware, Gemma models do not necessitate quantization and feature an impressive context length of 8,000 tokens [65] and Gemma exceeds the performance of larger models on important benchmarks, all while maintaining our strict standards for producing safe and responsible outputs [66].

This is why we choose gemma for predicting AMPs.

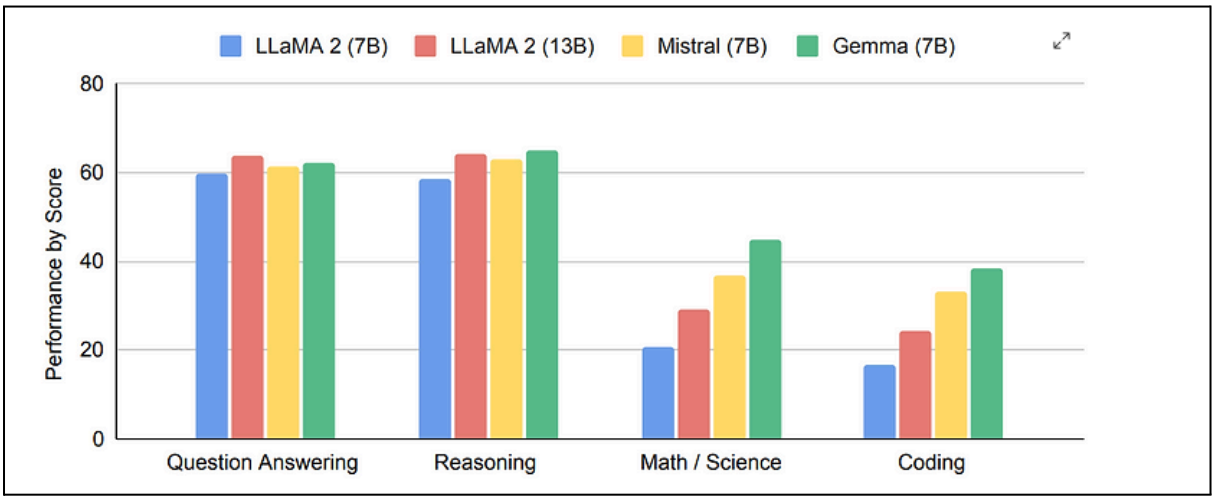

Figure 04: Gemma's Performance vs other LLMs in question Answering, Reasoning, Math/Science and Coding [65]

1.2. Antimicrobial peptides (AMPs) datasets

In our work the large language model gemma 2B is fine-tuned on a dataset for prediction of AMPs or non-AMPs. Our dataset originates from [52] .

This dataset curated by [67] was used to fine–tune our AMP-Gemma model. This dataset included 1778 positive samples consisting of screened sequences from the APD3 database (version 3) [68], and 1778 negative samples consisting of sequences from UniProt.

The data set used for the test also originates from [52] and includes 3973 antimicrobial peptide sequences: 2065 AMPs and 1908 non-AMPs. Screened positive samples from DRAMP database [69] and screened negative samples from AmPEP.

1.3. The question answering dataset

We have compiled a comprehensive dataset comprising specific and general information about antimicrobial peptides (AMPs), including their mechanisms of action, classifications, chemical characteristics, and structural properties that impact their functions. These data were curated from scholarly articles. This dataset consists of 143 questions and information pertinent to antimicrobial peptides was used to fine-tune our model for the Question Answering task of our AMP-Gemma.

1.4. High performance computing (HPC)

High Performance Computing (HPC) involves aggregating computing power to tackle complex problems in science, engineering, and business. The goal of HPC is to speed up computer programs and enhance overall work efficiency [67].

High Processing Center	**Characteristics**
CPU	12 nodes (2*14 cores),
GPU	2 nodes (4 GPU Tesla V100)
RAM	128 GB per node

Table 02: Basic information on High Processing Center (HPC) Characteristics that was used for data preprocessing and model training.

1.5. Python

Python, High Level ,General purpose programming language and one of the rapidly growing programming language today, offers several advantages for biologists and scientists:

- Widely embraced in the scientific community.
- Boasts well-crafted libraries for intricate scientific computations.
- Easily integrates with existing tools.
- Equipped with features ideal for manipulating strings, such as DNA bases and protein residues, which are of great interest to biologists.
- Used in Different fields like Data science, Big Data, and Artificial Intelligence.

1.6. GNU/linux

GNU/Linux, a free Open Source operating system, stands at the forefront of the Open Source community, bringing together numerous projects to successful fruition. When utilizing a GNU/Linux operating system, we engage with a suite of utility software primarily stemming from the GNU project built upon the Linux kernel. Hence, the system essentially comprises GNU with a Linux kernel. The expansive realm of GNU/Linux transcends specific companies or communities, offering the flexibility for individuals to tailor their own system to meet their unique requirements [69].

1.7. Packages

1.7.1. PyTorch

PyTorch stands out due to its unique design choices like exposed tensor strides, common aliasing views, and the ability to mutate data and metadata in place. Adapting PyTorch programs to a different compiler poses significant challenges due to these distinctive features [70].

PyTorch, an open-source machine learning framework, is built on the Python programming language and the Torch library. Torch, written in Lua, is another open-source ML library focused on deep neural networks. PyTorch is a top choice for deep learning research, designed to streamline the transition from research prototyping to deployment [71].

1.7.2. Transformers

Is State-of-the-art Machine Learning for JAX, PyTorch and TensorFlow. Transformers offers numerous pretrained models for tasks across text, image, and audio modalities. These tasks range from text classification and image segmentation to speech recognition and video classification. Additionally, Transformer models can handle combined modalities for tasks like table question answering, optical character recognition, and visual question answering [72].

1.7.3. Pandas

Pandas is a BSD-licensed open-source library that offers high-performance data structures and data analysis tools for the Python programming language. It is known for being easy to use [73].

1.7.4. NumPy

NumPy is the essential package for scientific computing in Python [74], it offers:

- a robust N-dimensional array object
- advanced (broadcasting) functions
- utilities for incorporating C/C++ and Fortran code
- valuable capabilities for linear algebra, Fourier transform, and random number generation.

1.7.5. PEFT

State-of-the-art Parameter-Efficient Fine-Tuning (PEFT) methods address the challenge of fine-tuning large pre-trained models, which can be prohibitively expensive due to their size. These methods allow for the effective adaptation of such models to different downstream tasks by fine-tuning only a small subset of additional model parameters, rather than all parameters. This approach leads to substantial reductions in computational and storage requirements. The recent state-of-the-art PEFT techniques have demonstrated performance levels on par with fully fine-tuned models [75].

1.7.6. LoRA

Low-Rank Adaptation (LoRA) involves freezing the weights of a pre-trained model and introducing trainable rank decomposition matrices into each layer of the Transformer architecture. This method significantly decreases the number of trainable parameters for downstream tasks. LoRA enables the training of certain dense layers indirectly by optimizing the rank decomposition matrices instead of adjusting the pre-trained weights [76].

1.7.7. Matplotlib

Matplotlib is a versatile library that allows for the creation of static, animated, and interactive visualizations in Python. It generates high-quality figures in multiple formats and environments, including hardcopy and interactive platforms. Matplotlib is compatible with Python scripts, Python/IPython shells, web application servers, and various graphical user interface toolkits [77].

1.7.8. Django

Django is a Python web framework that promotes rapid development and clean, pragmatic design [78] .

Packages	**Version**
PyTorch	PyTorch 2.2.1
Transformers	Transformers 4.38.2
Pandas	Pandas 2.2.1
NumPy	NumPy 1.24.4
PEFT	PEFT 0.9.0
Matplotlib	Matplotlib 3.8.4
Django	Django 5.0.6

Table 03 : The versions of the used packages.

2. Methods

In this section, we provide a detailed and concise description of the experimental procedures chosen, along with the various protocols utilized throughout the steps of the Large language model process. These procedures are presented in chronological order to effectively demonstrate how they were applied to achieve the objectives outlined in this document.

2.1. Data preparation

2.1.1. Importing the necessary libraries

We start with importing the necessary libraries for data manipulation, numerical operations :

- numpy: For numerical operations involving arrays and matrices.
- pandas: For data manipulation and analysis.
- train_test_split: To split the dataset into training and testing sets.
- StandardScaler: For feature scaling.
- accuracy_score and classification_report: For evaluating model performance.
- GemmaModel: From the GEMMA library, used to implement the Generative Encoder Model for Molecular Attention.

2.1.2. Load data

2.1.2.1. For AMPs prediction task

Following the library imports, the next step involves loading the train dataset and the test set from [52], the train dataset is contained within a CSV file comprising 3556 antimicrobial peptide sequences, the sequences are methodically classified into 1778 antimicrobial peptides (AMPs) and 1778 non-AMPs, with each sequence paired with its specific antimicrobial activity label. The test set includes 2065 AMPs and 1908 non-AMPs.

We curated a test dataset for external model validation by combining the train and the test set, the data validation file. Then we split this data randomly in two parts: train and test part : 80% for train and 20% for test.

The CSV file 'amp_data.csv' was imported into a Pandas data frame named 'df' in Python. This data frame encompasses specific columns for peptide sequences and their respective labels that denote antimicrobial activity.

2.1.2.2. For Questions answering task

For the second task of our model, we curated a dataset comprising 149 inquiries and data pertaining to antimicrobial peptides in the form of a CSV file named 'QA_AMPs.csv'. The CSV file consists of two columns, with the primary column dedicated to the questions and the secondary column for the corresponding answers. Subsequently, we transformed these two columns into prompts structured into two components: "instruction" and "response", with the question allocated to the instruction segment and the answer assigned to the response segment.

2.2. Data Visualization

Data visualization is a crucial aspect that helps identify key features within the data, shedding light on both antimicrobial and non-antimicrobial peptide sequences in the training and test sets. This process is vital in increasing understanding of data exploration, interpretation, and communication in antimicrobial peptide prediction tasks. Furthermore, it assists in comprehending how the model generates predictions, ultimately enhancing the accuracy of such predictions.

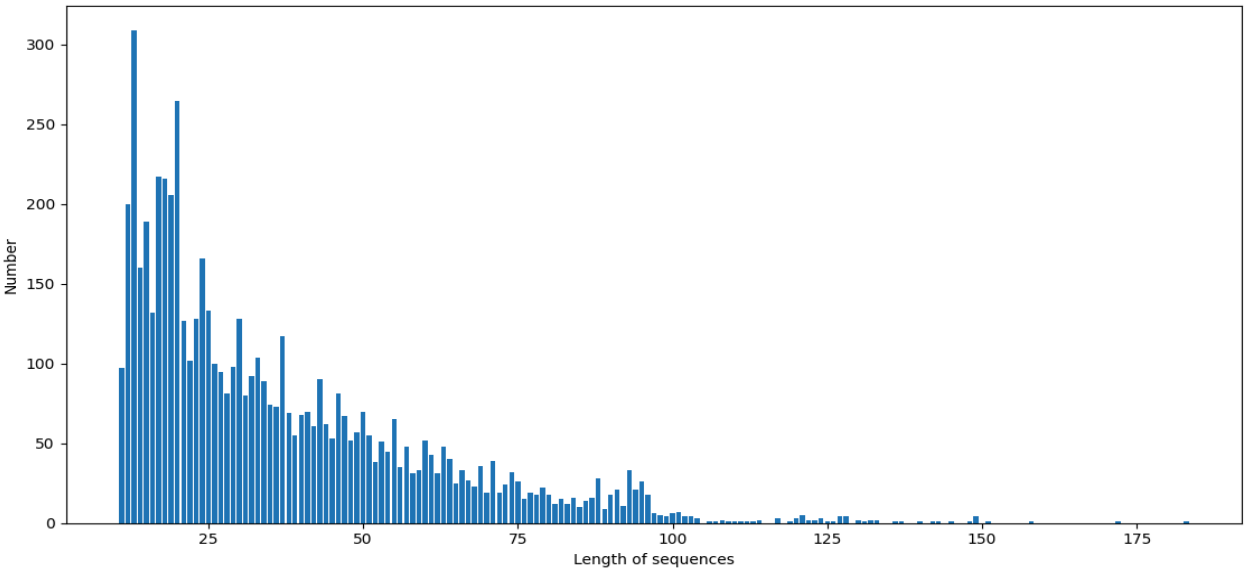

Figure 05: Amino acid length distributions of the antimicrobial peptides (AMPs) and Non-AMPs from the fine-tuning set.

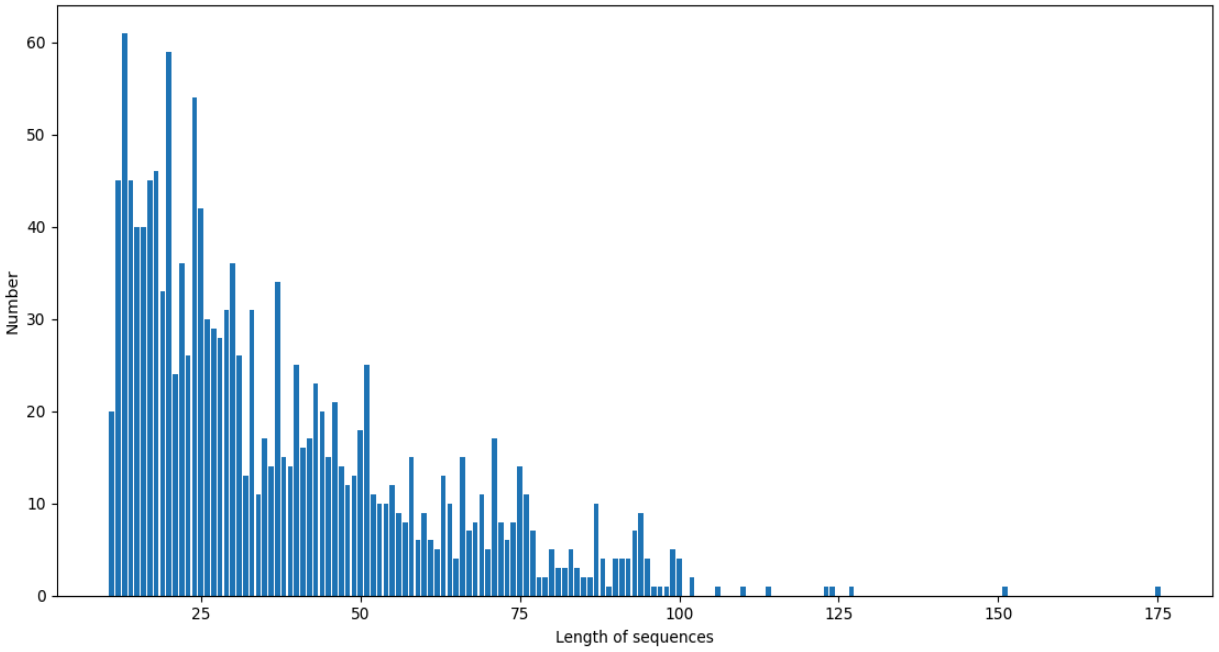

Figure 06 : Amino acid length distributions of the AMPs and the non-AMPs from the test set.

2.3. Base model

Next, we load in our base model the Gemma large language model, the pretrained GEMMA 2B model is available on :

- Instruction fine-tuned version of the base 2B model: https://huggingface.co/google/gemma-2b-it

2.4. Preprocess data

In this step, we need to preprocess our data so that it can be used for training. This consists of using a tokenizer to convert the text into an integer representation understood by the base model.

2.4.1. For AMPs prediction task

In our first case GEMMA requires the peptide sequences of the train set to be tokenized into a format that the model can process. This involves converting each amino acid in the sequences into a numerical token.

To apply the tokenizer to the dataset, we utilize a class from the transformers library called AutoTokenizer. We then employ the from_pretrained() method to load a pre-trained tokenizer file, such as "tokenizer.model", specifying a maximum sequence length of 200. Subsequently, we iterate through each sequence in the training data, splitting it into an array of characters, and convert each character into a numerical representation.

```
self.tokenizer = AutoTokenizer.from_pretrained(tokenizer_name, do_lower_case=False)
```

Figure 07 : Applying tokenizer to the dataset.

2.4.2. For question answering task

In this second case, GEMMA requires each word in the second dataset to be tokenized into a format that is readable by the model. This process includes dividing the input data into characters, words, and subwords, and then converting it into numerical lists, arrays, or matrices.

2.5. Training the model

We train the GEMMA model on the preprocessed training data. However, fine-tuning a pre-trained GEMMA model can significantly improve its performance for the specific task of AMP prediction and the question answering task.

2.5.1. Untrained model performance

2.5.1.1. for AMPs prediction task

Before training our model, we initialize the GEMMA model with pre-trained weights so that we can evaluate how the base model with a randomly initialized classification head performs on some example inputs.

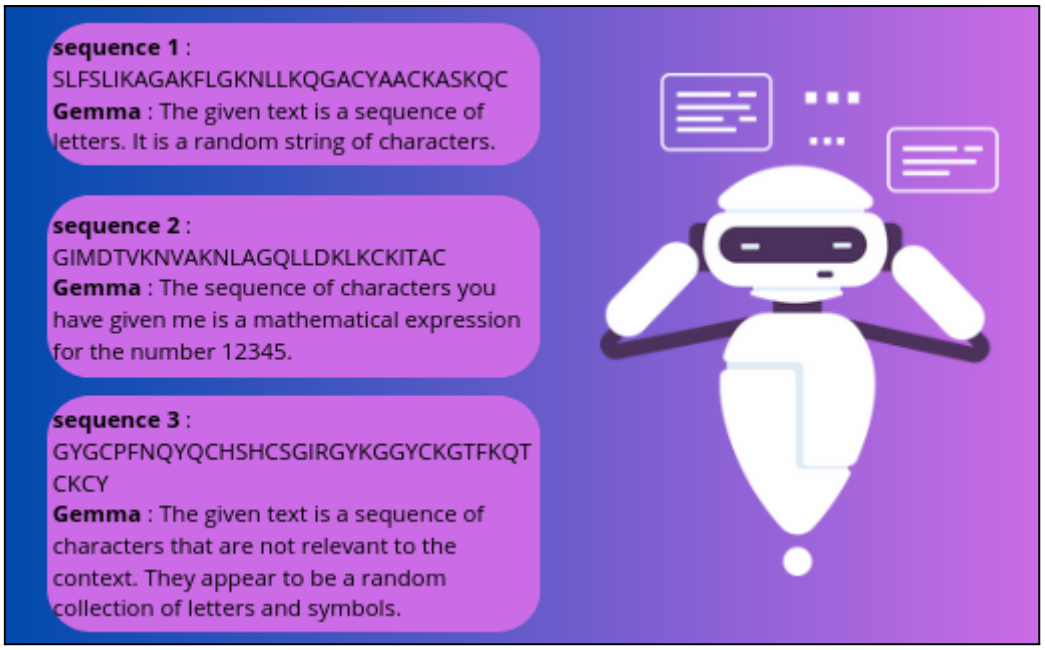

Figure 08 : AMPs prediction results before fine-tuning.

2.5.1.2. for question answering task

Before training our model for the question-answering task, we gave our chatbot a random question. The responses generated by the chatbot were general, average, and vague, as depicted in figure 09.

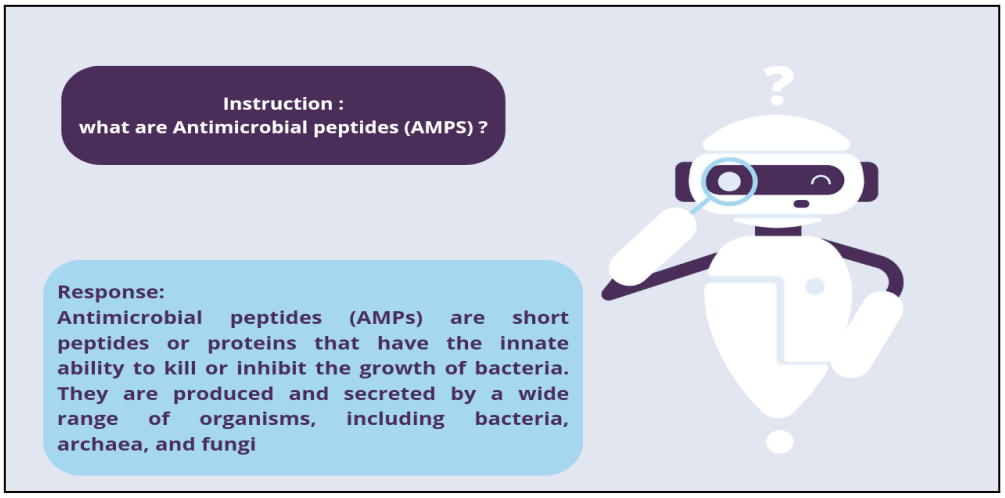

Figure 09 : Results of question answering task before the fine-tuning.

As expected, the model performance in both tasks, AMPs classification and question answering is equivalent to random guessing. Let's see how we can improve this with fine-tuning.

2.6. Fine-tuning the model

2.6.1. for AMPs prediction task

We fine-tuned the model for this task using lora. To utilize LoRA for fine-tuning, a configuration file is required initially. This file specifies all the parameters for the LoRA algorithm. Subsequently, a new iteration of our model can be generated, which is compatible with training through PEFT. Following that, we establish the hyperparameters for model training.

For identifying antimicrobial peptides in this work, we fine-tune the Gemma 2b Large Language Model to distinguish between AMPs and non-AMPs. We set the number of epochs, batch size and learning rate multiplier during training to 15, 16, 5×10^{-5} and respectively.

2.6.2. for Question Answering task

We fine-tuned the model for this task using the "keras_nlp" library. Specifically, we utilized a class named "GemmaCausalLM" which enabled us to leverage the Gemma model "gemma_1.1.instruct_2b_en". After setting the rank of adaptation to 4 (r = 4) with the call backbone.enable_lora(), we employed the optimizer method "AdamW" from the "keras" library. The learning rate was set to 5e-5, the batch size to 1, and the number of epochs to 20 before proceeding with the training.

2.7. Evaluating the model

2.7.1. for AMPs prediction task

Here, we compared the binary class prediction performance of our fine-tuned model using the curated test set to measure the performance of our classification model. We curated this test dataset for external model validation by combining the train and the test set from [52], then we split this data into two parts: 80% for training and 20% for testing. After that, we used evaluation metrics to compute the model's accuracy and evaluated our fine-tuned model.

Figure 10: AMPs prediction results after fine tuning.

2.7.2. for question answering task

After fine-tuning our model for the question-answering task, we gave our chatbot another random question. This time the responses generated by the model chatbot were exceptional and specific, as shown in the figure 11.

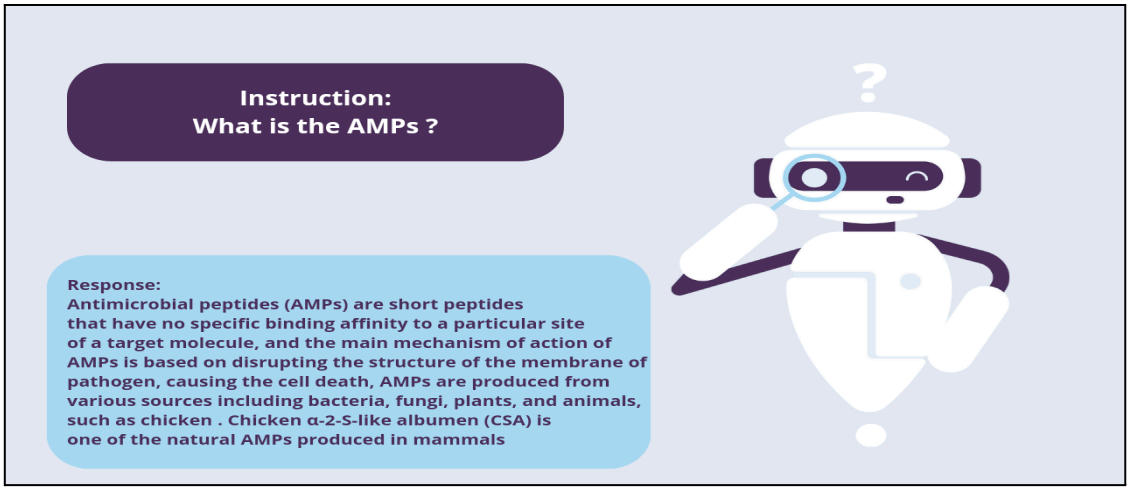

Figure 11: Results of question answering task after the fine-tuning.

2.8. The AMP-Gemma web application

In order to provide a user-friendly experience, we designed a user interface that allows users to easily input their data and access the results generated by our model. The platform also includes interactive features such as visualization tools, customizable settings, and easy-to-understand explanations of the model's outputs. Additionally, we have integrated tutorials and customer support to assist users in navigating the platform and making the most of the model's capabilities. Our goal is to make the model accessible and practical for users of all levels of expertise.

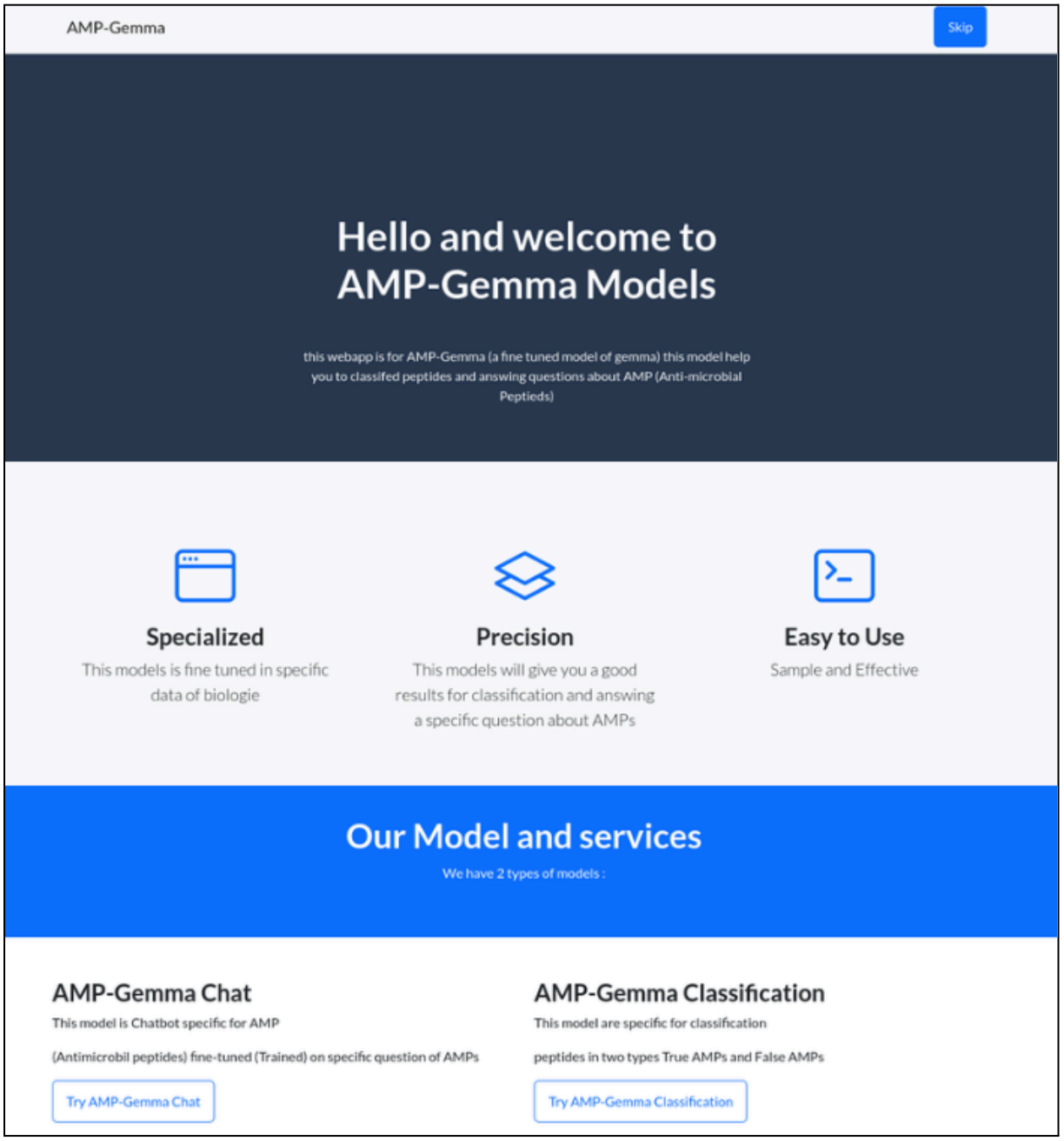

Figure 12: The AMP-Gemma web application.

CHAPTER III :

Results and Discussion

1. Results

1.1. Identifying antimicrobial peptides

We perform the fine-tuning stage. The Gemma 2b model is tested on the test set and the results for our model are shown in Table 4.

The accuracy achieved on test samples was 0.8624. As shown, we obtained a very good accuracy in all experiments. Besides the accuracy of the model, The precision obtained on the test samples was 0.8766, the recall and F1 score was 0.8529 and 0.8646 respectively. The obtained results for all metrics are shown in table 3.

	Recall	**ACC**	**F1**	**Precision**
test-set	0.8529	0.8624	0.8646	0.8766

table 04 : The Results of the AMP-GEMMA for identifying the antimicrobial peptides on the test and train set.

1.1.1. Evaluation metrics

The binary classification performance of our fine-tuned model for AMP prediction, AMP-Gemma, was evaluated using six widely accepted metrics to assess the performance, namely, sensitivity/Recall (SN), F1-score (F1), accuracy (ACC),and Precision.

The aforementioned metrics were calculated using the four different types of prediction outcomes: true positive (TP), false positive (FP), true negative (TN), and false negative (FN).

1.1.1.1. Precision

Precision measures the number of correct instances retrieved divided by all retrieved instances .

$$Precision = P = \frac{TP}{TP+FP}$$

1.1.1.2. Recall

Sensitivity / True positive rate (TPR)/recall, is an evaluation concept used in medicine and health informatics .

measures the proportion of negatives that are correctly identified (e.g., the percentage of healthy people who are correctly identified as not having the condition).

$$SN = TPR = \frac{TP}{TP+FN}$$

1.1.1.3. Accuracy

is another measurement defined as the proportion of true instances retrieved, both positive and negative, among all instances retrieved. Accuracy is a weighted arithmetic mean of precision and inverse precision. Accuracy can also be high but precision low, meaning the system performs well but the results produced are slightly spread, compare this with hitting the bulls eye meaning both high accuracy and high precision.

$$Accuracy: A = \frac{TP+TN}{TP+TN+FP+FN}$$

1.1.1.4. F1 Score

The F-score is defined as the weighted average of both precision and recall depending on the weight function β, The F1-score means the harmonic mean between precision and recall, The F-score is also called the F-measure. The F1-score can have different indices giving different weights to precision and recall.

$$F1 = F = \frac{2TP}{2TP+FP+FN}$$

1.2. Question answering task

For this second task of our model, we fine-tune the Gemma 2b Large Language Model to generate specific answers about the antimicrobial peptides. We set the number of epochs, batch size and learning rate multiplier during training to 15, 1 and 5×10^{-5} respectively.

2. Discussion

2.1. AMP-Gemma performance

2.1.1. Identifying antimicrobial peptides

In the first task of differentiating between antimicrobial peptides (AMPs) and non-AMPs, our fine-tuned large language model, AMP-GEMMA, has exhibited excellent performance. It has demonstrated superior accuracy compared to other conventional machine learning and deep learning approaches, as evidenced in Table 5.

The impressive outcomes achieved by large language models, such as our AMP-GEMMA and GPT-3.5 (davinci) model [4] in predicting AMPs, highlight the potential of these models to outperform traditional machine learning and deep learning methods in tasks related to AMP identification.

The AMP-GEMMA model has displayed higher accuracy in AMP prediction than the GPT-3.5 (davinci) model. Specifically, the accuracy of AMP-GEMMA was measured at 86.24%, surpassing the 71.8% accuracy of GPT-3.5. These significant findings underscore the efficacy of our model in accurately predicting antimicrobial peptides.

To access the performance of our fine-tuned model Gemma 2b, we conduct a comparison with the advanced protein large language model, ESM (esm msa1b t12 100M UR50S), several machine learning based methods, i.e., XGBoost (XGB), Multinomial Naive Bayes (MNB), Support Vector Machines (SVM), KNearest Neighbor (KNN), Logistic Regression (LR), MultiLayer Perceptron (MLP), Random Forest (RF) , GBoost (GB)[81] , and AMP-BERT and GPT-3.5 (Davinci-ft) using two datasets from [52].

Model	**Recall**	**ACC**	**F1**
XGB	0.695	0.660	0.654
MNB	0.687	0.711	0.746
SVM	0.740	0.706	0.702
KNN	0.608	0.632	0.698
LR	0.724	0.699	0.702
MLP	0.701	0.707	0.730
RF	0.714	0.703	0.715
GB	0.708	0.646	0.616
ESM	0.865	0.742	0.688
AMP-BERT	0.876	0.792	0.760
GPT-3.5 (Davinci-ft)	0.844	0.718	0.759
AMP-Gemma	0.8529	0.8624	0.8646

table 5: The compared results of different models for identifying antimicrobial peptides on the test set.

2.1.2. Question answering task

In the second stage of our model's analysis, the pre-fine-tuning performance in answering questions was characterized by a lack of specificity. Following the fine-tuning process, there was a notable enhancement in the performance and precision of our model AMP-Gemma , leading to a significant improvement in accuracy for text generation tasks. The results have been visualized in Figure 11. This successful outcome can be attributed to the utilization of a powerful large language model that was fine-tuned using targeted data related to the properties, functions, and structural characteristics of antimicrobial peptides.

Conclusion

Conclusion

In conclusion, the applications of the Large Language Models in bioinformatics research are evaluated in two basic tasks, including identifying antimicrobial peptides and question answering. The study demonstrates that LLMs, like our model the AMP-Gemma, can achieve remarkable performance on these tasks with proper prompts and models.

With the rapid growth in related knowledge and lead compounds, more AMPs may enter clinical tests and treatment in the near future. Our proposed model for AMP prediction demonstrates the value of large language models for AMP-based drug candidates, which can accelerate the discovery process of antimicrobial drugs.

Due to its ability to be generalized to unseen data and learn meaningful features of peptides, AMP-Gemma is expected to contribute to the development of novel antimicrobial drugs against which microbes cannot easily evolve resistance. As more AMPs are experimentally validated and new structural information is uncovered, data-driven computational methods, such as our model, will become even more effective and greatly contribute to the field of AMP-based drug discovery. Additionally, our AMP-Gemma chatbot can guide these experiments by offering specific and specialized information about antimicrobial peptides, facilitating drug discovery for the scientific community in the AMP field.

As a future direction for our work, we aim to achieve the following goals:

- This work is expected to facilitate researchers in bioinformatics by providing guidance on the use of advanced LLMs, there by promoting the development of AI for scientific applications.

- We intend to refine our model for predicting other peptide families and improve its applicability for the drug discovery process.

- We plan to evaluate the performance of various large language models across diverse tasks in our field.

- Lastly, we aim to explore the use of LLMs in other bioinformatics problems.

Références Bibliographiques

[1]: Luong, H. X., Thanh, T. T., & Tran, T. H. (2020). Antimicrobial peptides – Advances in development of therapeutic applications. *Life Sciences, 260*, 118407. https://doi.org/10.1016/j.lfs.2020.118407

[2]: World Health Organization: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance, Visited on (2023, November 21).

[3]: Waghu, F. H., Barai, R. S., Gurung, P., & Idicula-Thomas, S. (2016). CAMPR3: A database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Research, 44*(D1), D1094–D1097. https://doi.org/10.1093/nar/gkv1051

[4]: Yin, H., Gu, Z., Wang, F., Yiparemu, Zhu, Y., Tu, X., Hua, X.-S., Luo, X., & Sun, Y. (2024, February 21). An evaluation of large language models in bioinformatics research.

[5]: Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial peptides: Classification, design, application and research progress in multiple fields. *Frontiers in Microbiology, 11*, 582779.https://doi.org/10.3389/fmicb.2020.582779

[6]: Khaldi, N. (2012). Bioinformatics approaches for identifying new therapeutic bioactive peptides in food. *Functional Foods in Health and Disease, 2*(10), 325-338.

[7]: Zhong, Z., Zhou, K., & Mottin, D. (2024). Benchmarking large language models for molecule prediction tasks. *arXiv preprint* arXiv:2403.05075.

[8]: Hendy, A., Abdelrehim, M., Sharaf, A., Raunak, V., Gabr, M., Matsushita, H., Kim, Y. J., Afify, M., & Awadalla, H. H. (2023). How good are GPT models at machine translation? A comprehensive evaluation. *arXiv preprint* arXiv:2302.09210.

[9]: Krishna, S., Ma, J., Slack, D., Ghandeharioun, A., Singh, S., & Lakkaraju, H. (2023). Post hoc explanations of language models can improve language models. *arXiv preprint* arXiv:2305.11426.

[10]: Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. (2023). Sparks of artificial general intelligence: Early experiments with GPT-4. *arXiv preprint* arXiv:2303.12712.

[11]: Chernyavskiy, A., Ilvovsky, D., & Nakov, P. (2021). Transformers: “The end of history” for natural language processing? In *Machine Learning and Knowledge Discovery in Databases. Research Track: European Conference, ECML PKDD 2021, Bilbao, Spain, September 13–17, 2021, Proceedings, Part III* (pp. 677–693). Springer.

[12]: Naveed, H., Khan, A. U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., & Mian, A. (2024, February 20). A comprehensive overview of large language models.

[13]: Guntuboina, C., Das, A., Mollaei, P., Kim, S., & Farimani, A. B. (2023). PeptideBERT: A language model based on transformers for peptide property prediction. *The Journal of Physical Chemistry Letters*.

[14]: Bahar, A. A., & Ren, D. (2013). Antimicrobial peptides. *Pharmaceuticals, 6*(12), 1543-1575. https://doi.org/10.3390/ph6121543

[15]: Wang, G. (2023). The antimicrobial peptide database is 20 years old: Recent developments and future directions. *The Protein Society*.

[16]: Sadredinamin, M., Mehrnejad, F., Hosseini, P., & Doustdar, F. (2015). Antimicrobial peptides (AMPs). *Infectious Diseases and Tropical Medicine Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.* Accepted: 29 December 2015.

[17]: Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial peptides: Classification, design, application and research progress in multiple fields. *Frontiers in Microbiology, 11*, 582779. https://doi.org/10.3389/fmicb.2020.582779

[18]: Bastian, A., & Schafer, H. (2001). Human alpha-defensin 1 (hnp-1) inhibits adenoviral infection in vitro. *Regulatory Peptides, 101*, 157-161.

[19]: Horne, W. S., Wiethoff, C. M., Cui, C., Wilcoxen, K. M., Amorin, M., Ghadiri, M. R., & Nemerow, G. R. (2005). Antiviral cyclic D,L-α-peptides: Targeting a general biochemical pathway in virus infections. *Bioorganic & Medicinal Chemistry, 13*, 5145-5153.

[20]: Robinson, W. E., Jr., McDougall, B., Tran, D., & Selsted, M. E. (1998). Anti-HIV-1 activity of indolicidin, an antimicrobial peptide from neutrophils. *Journal of Leukocyte Biology, 63*, 94-100.

[21]: Sitaram, N., & Nagaraj, R. (1999). Interaction of antimicrobial peptides with biological and model membranes: Structural and charge requirements for activity. *Biochimica et Biophysica Acta, 1462*, 29-54.

[22]: Shai, Y. (2002). Mode of action of membrane active antimicrobial peptides. *Biopolymers, 66,* 236-248.

[23]: Zhang, L., Rozek, A., & Hancock, R. E. (2001). Interaction of cationic antimicrobial peptides with model membranes. *Journal of Biological Chemistry, 276*, 35714-35722.

[24]: Jenssen, H., Hamill, P., & Hancock, R. E. (2006). Peptide antimicrobial agents. *Clinical Microbiology Reviews, 19*, 491-511.

[25]: De Lucca, A. J., Bland, J. M., Jacks, T. J., Grimm, C., & Walsh, T. J. (1998). Fungicidal and binding properties of the natural peptides cecropin B and dermaseptin. *Medical Mycology, 36*, 291-298.

[26]: De Lucca, A. J., & Walsh, T. J. (1999). Antifungal peptides: Novel therapeutic compounds against emerging pathogens. *Antimicrobial Agents and Chemotherapy, 43,* 1-11.

[27]: Lee, Y. T., Kim, D. H., Suh, J. Y., Chung, J. H., Lee, B. L., Lee, Y., & Choi, S. (1999). Structural characteristics of tenecin 3, an insect antifungal protein. *Biochemical and Molecular Biology International, 47*, 369-376.

[28]: Lehrer, R. I., Szklarek, D., Ganz, T., & Selsted, M. E. (1985). Correlation of binding of rabbit granulocyte peptides to Candida albicans with candidacidal activity. *Infection and Immunity, 49*, 207-211.

[29]: Terras, F. R., Schoofs, H. M., De Bolle, M. F., Van Leuven, F., Rees, S. B., Vanderleyden, J., Cammue, B. P., & Broekaert, W. F. (1992). Analysis of two novel classes of plant antifungal proteins from radish (Raphanus sativus L.) seeds. *Journal of Biological Chemistry, 267*, 15301-15309.

[30]: Van der Weerden, N. L., Hancock, R. E., & Anderson, M. A. (2010). Permeabilization of fungal hyphae by the plant defensin nad1 occurs through a cell wall-dependent process. *Journal of Biological Chemistry, 285*, 37513-37520.

[31]: Moerman, L., Bosteels, S., Noppe, W., Willems, J., Clynen, E., Schoofs, L., Thevissen, K., Tytgat, J., Van Eldere, J., & van der Walt, J. (2002). Antibacterial and antifungal properties of α-helical, cationic peptides in the venom of scorpions from southern Africa. *European Journal of Biochemistry, 269*, 4799-4810.

[32]: Zasloff, M. (1987). Magainins, a class of antimicrobial peptides from Xenopus skin: Isolation, characterization of two active forms, and partial cDNA sequence of a precursor. *Proceedings of the National Academy of Sciences of the USA, 84*, 5449-5453.

[33]: Li, Q., Zhou, W., Wang, D., Wang, S., & Li, Q. (2020). Prediction of anticancer peptides using a low-dimensional feature model. *Frontiers in Bioengineering and Biotechnology, 8*, 892. https://doi.org/10.3389/fbioe.2020.00892

[34]: Sun, Y., Zhang, W., Chen, Y., Ma, Q., Wei, J., & Liu, Q. (2016). Identifying anti-cancer drug response related genes using an integrative analysis of transcriptomic and genomic variations with cell line-based drug perturbations. *Oncotarget, 7*, 9404.

[35]: Liu, H., Luo, L. B., Cheng, Z. Z., Sun, J. J., Guan, J. H., Zheng, J., ... & Rao, A. R. (2018). Group-sparse modeling drug-kinase networks for predicting combinatorial drug sensitivity in cancer cells. *Current Bioinformatics, 13*, 437-443. https://doi.org/10.2174/1574893613666180118104250

[36]: Wu, D., Gao, Y., Qi, Y., Chen, L., Ma, Y., & Li, Y. (2014). Peptide-based cancer therapy: Opportunity and challenge. *Cancer Letters, 351*, 13-22. https://doi.org/10.1016/j.canlet.2014.05.002

[37]: Gaspar, D., Veiga, A. S., & Castanho, M. A. R. B. (2013). From antimicrobial to anticancer peptides. *Frontiers in Microbiology, 4*, 294. https://doi.org/10.3389/fmicb.2013.00294

[38]: Gordon, Y., & Eric, G. (2005). A review of antimicrobial peptides and their therapeutic potential as anti-infective drugs. *Current Eye Research, 30*(7), 505-515.

[39]: de la Fuente-Núñez, C., Silva, O. N., Lu, T. K., & Franco, O. L. (2017). Antimicrobial peptides: Role in human disease and potential as immunotherapies. *Pharmacology & Therapeutics, 178*, 132-140. https://doi.org/10.1016/j.pharmthera.2017.04.002

[40]: Rai, M., Pandit, R., Gaikwad, S., & Kövics, G. (2016). Antimicrobial peptides as natural bio-preservatives to enhance the shelf-life of food. *Journal of Food Science and Technology, 53*, 3381-3394. https://doi.org/10.1007/s13197-016-2318-2315

[41]: Santos, J. C. P., Sousa, R. C. S., Otoni, C. G., Moraes, A. R. F., Souza, V. G. L., Medeiros, E. A. A., ... & Dallago, C. (2018). Nisin and other antimicrobial peptides: Production, mechanisms of action, and application in active food packaging. *Innovative Food Science & Emerging Technologies, 48*, 179-194. https://doi.org/10.1016/j.ifset.2018.06.008

[42]: Fox, J. L. (2013). Antimicrobial peptides stage a comeback. *Nature Biotechnology, 31*(5), 379-382.

[43]: Hancock, R. E., & Scott, M. G. (2000). The role of antimicrobial peptides in animal defenses. *Proceedings of the National Academy of Sciences of the USA, 97*, 8856-8861.

[44]: Robinson, W. E., Jr., McDougall, B., Tran, D., & Selsted, M. E. (1998). Anti-HIV-1 activity of indolicidin, an antimicrobial peptide from neutrophils. *Journal of Leukocyte Biology, 63*, 94-100.

[45]: Selsted, M.E., Novotny, M.J., Morris, W.L., Tang, Y.Q., Smith, W., & Cullor, J.S. (1992). Indolicidin, a novel bactericidal tridecapeptide amide from neutrophils. *Journal of Biological Chemistry, 267*(16), 4292–4295.

[46]: Tiwari, T., Tiwari, T., & Tiwari, S. (2018). How Artificial Intelligence, Machine Learning and Deep Learning are Radically Different? *International Journal of Advanced Research in Computer Science and Software Engineering, 8*(2), ISSN: 2277-128X.

[47]: Sadrehaghighi, I. (n.d.). Artificial Intelligence (AI) & Machine Learning (ML/DL/NNs). A N N A P O L I S, MD.

[48]: McCarthy, J. (2007). What is artificial intelligence.

[49]: Kühl, N., Schemmer, M., Goutier, M., & Satzger, G. (2022). Artificial intelligence and machine learning. *Electronic Markets, 32*(4), 2235-2244.

[50]: Kaggle. (Visited 4 April 2024) . https://www.kaggle.com/discussions/general/381324

[51]: Kim, S., Mollaei, P., Antony, A., Magar, R., & Farimani, A. B. (2023). GPCR-BERT: Interpreting Sequential Design of G Protein Coupled Receptors Using Protein Language Models. *arXiv*. DOI: 10.48550/arXiv.2310.19915.

[52]: Lee, J. (2022). AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model. *Protein Science*.

[53]: Moretta, A., Scieuzo, C., Petrone, A. M., Salvia, R., Manniello, M. D., Franco, A., ... Falabella, P. (2021). Antimicrobial peptides: A new hope in biomedical and pharmaceutical fields. *Frontiers in Cellular and Infection Microbiology, 11*, 668632.

[54]: Meher, P. K., Sahu, T. K., Saini, V., & Rao, A. R. (2017). Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Scientific Reports, 7*, 1–12.

[55]: Bhadra, P., Yan, J., Li, J., Fong, S., & Siu, S. W. (2018). AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific Reports, 8*, 1–10.

[56]: Lata, S., Mishra, N. K., & Raghava, G. P. (2010). AntiBP2: Improved version of antibacterial peptide prediction. *BMC Bioinformatics, 11,* 1–7.

[57]: Porto, W. F., Pires, A. S., & Franco, O. L. (2012). CS-AMPPred: An updated SVM model for antimicrobial activity prediction in cysteine-stabilized peptides. *PLoS One, 7*, e51444.

[58]: Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. *Bioinformatics, 34*, 2740–7.

[59]: Su, X., Xu, J., Yin, Y., Quan, X., & Zhang, H. (2019). Antimicrobial peptide identification using multi-scale convolutional network. *BMC Bioinformatics, 20*, 1–10.

[60]: Fu H, Cao Z, Li M, Wang S. ACEP: improving antimicrobial peptides recognition through automatic feature fusion and amino acid embedding. *BMC Genomics*. 2020;21:1–14.

[61]: Elnaggar, A.; Heinzinger, M.; Dallago, C.; Rehawi, G.; Wang, Y.; Jones, L.; Gibbs, T.; Feher, T.; Angerer, C.; Steinegger, M.; et al. Prottrans:

Toward Understanding the Language of Life Through SelfSupervised Learning. IEEE PAMI 2022, 44, 7112–7127.

[62]: AgeMagician. Visited April 8, 2024. ProtTrans. GitHub. https://github.com/agemagician/ProtTrans

[63]: Rostlab. Visited April 12, 2024. ProtBERT. Hugging Face. https://huggingface.co/Rostlab/prot_bert

[64]: Google Cloud. Retrieved April 15, 2024. Performance Deepdive of GEMMA on Google Cloud. https://cloud.google.com/blog/products/ai-machine-learning/performance-deepdive-of-gemma-on-google-cloud?hl=en

[65]: The AI Dream. Visited March 3, 2024. Google GEMMA: Open-source LLM - Everything You Need to Know. https://www.theaidream.com/post/google-gemma-open-source-llm-everything-you-need-to-know

[66]:GEMMA. Open Models , https://blog.google/technology/developers/gemma-open-models/, Visited on (1 March 2024).

[67]: Chen, S. Introduction to High Performance Computing. Research Computing Services (RCS) Boston University.

[68]: Guo, G., Wang, H., Bell, D., Bi, Y., & Greer, K. (2003). Knn model-based approach in classification. In *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE: OTM Confederated International Conferences, CoopIS, DOA, and ODBASE 2003, Catania, Sicily, Italy, November 3-7, 2003. Proceedings* (pp. 986–996).

[69]: Jorba Esteve, J. (2009). Introduction to the GNU/Linux operating system. FUOC.GNUFDL • PID_00148470.

[70]: Ansel, J., et al. (2024). PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In *ASPLOS '24, April 27-May 1, 2024, La Jolla, CA, USA* (pp. 1–4). ACM. ISBN 979-8-4007-0385-0/24/04.

[71]: TechTarget: https://www.techtarget.com/searchenterpriseai/definition/PyTorch, Visited on (1 May 2024).

[72]: Transformers:https://pypi.org/project/transformers/ , Visited on (6 May 2024).

[73]: Pandas Documentation: https://pandas.pydata.org/docs/, Visited on (7 May 2024).

[74]: NumPy: https://pypi.org/project/numpy/1.24.4/ , Visited on (15 May 2024).

[75]: PEFT: https://pypi.org/project/peft/ , Visited on (15 May 2024).

[76]: Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., ... & Chen, W. (2021). Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

[77]: Matplotlib: https://pypi.org/project/matplotlib/, Visited on (16 May 2024).

[78]: Django: https://pypi.org/project/Django/ , Visited on (15 May 2024).

Academic year : 2023-2024	**Submitted by :** OUFEROUKH Oussema LAZOUNE Dalal

Prediction of antimicrobial peptide function based on a fine-tuned large language model

Dissertation Submitted in Partial Fulfillment of the Requirements for the Degree of Master in Bioinformatics

Abstract :

Antimicrobial peptides (AMPs) play a crucial role in the innate immune system by effectively combating disease-causing pathogens. The rapid increase of drug-resistant infections poses serious challenges to current antimicrobial therapies. Traditional wetlab experimentation for AMP identification is known to be cost-intensive. Therefore, the incorporation of efficient computational tools becomes imperative in the preemptive identification of optimal AMP candidates prior to in vitro experimentation.

In this study, we introduce AMP-Gemma, a fine-tuned large language model tailored for the precise prediction of antimicrobial peptides. Through our model, AMP-Gemma, we achieved an exceptional accuracy rate of 86.24% in accurately predicting peptides with antimicrobial activity, surpassing the performance of existing machine learning and deep learning methodologies for AMP prediction.

Furthermore, we have developed a user-friendly chatbot specializing in AMPs, designed to cater to the needs of the biological community. This innovative tool aims to facilitate access to information and streamline communication within the field of antimicrobial peptide research.

Keywords : Antimicrobial peptides (AMPs), AMP-Gemma, Large language models (LLMs), Machine learning, Deep learning, Antimicrobial activity.

President of Examiners :	Dr BOUHALOUF H (MC(B) - U Constantine 1 Frères Mentouri).
Supervisor :	Dr CHEHILI H (MC(A) - UFM Constantine 1).
Examiner :	Dr MEZIANI Y (MA(B) - UFM Constantine 1),