

**ALMA MATER STUDIORUM  
UNIVERSITÀ DI BOLOGNA**

---

**DEPARTMENT OF COMPUTER SCIENCE  
AND ENGINEERING**

ARTIFICIAL INTELLIGENCE

**MASTER THESIS**

in

Natural Language Processing

**AN LLM AGENT SYSTEM FOR REDUCING  
HALLUCINATIONS IN DRUG DISCOVERY  
TASKS**

CANDIDATE

Gabriele Fossi

SUPERVISOR

Prof. Paolo Torroni

CO-SUPERVISOR

Salvatore Raieli

Academic year 2023-2024

Session 5th

# Abstract

Artificial Intelligence has revolutionized drug discovery, significantly accelerating the process of identifying new therapeutics. Large Language Models (LLMs) can also play a crucial role in this process, but they are prone to hallucinations, generating inaccurate or misleading content. In this thesis, AI techniques such as Fine-Tuning and Retrieval-Augmented Generation (RAG) are employed in order to overcome this issue. Through the integration of RAG and other tools, an LLM Agent system is designed to retrieve information from external sources and generate an automatic target dossier in the context of pancreatic cancer. This system can significantly assist researchers by providing precise information on specific targets. The results show that fine-tuning and RAG enhance the LLM's expertise in the biomedical domain, resulting in more accurate and comprehensive responses. The Agent successfully generates an automatic target dossier with precise and relevant information in a minimal amount of time, showing its potential to streamline the drug discovery process. The integration of additional tools and models, along with the application of multimodal machine learning, might further improve the system, enabling it to generate a more informative and complete target dossier.

# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
<b>2</b>	<b>Background and Related Work</b>	<b>4</b>
2.1	AI need for Drug Discovery . . . . .	4
2.1.1	Pancreatic Ductal Adenocarcinoma . . . . .	4
2.1.2	The Drug Discovery Process . . . . .	6
2.1.3	AI in Drug Discovery . . . . .	8
2.2	Large Language Models in Medicine . . . . .	9
2.2.1	Transformers . . . . .	9
2.2.2	Large Language Models . . . . .	11
2.2.3	The Application of LLMs in Drug Discovery . . . . .	14
<b>3</b>	<b>Materials and Methods</b>	<b>16</b>
3.1	Data . . . . .	16
3.2	Model Fine-Tuning . . . . .	18
3.3	Retrieval-Augmented Generation . . . . .	20
3.4	LLM Agents . . . . .	24
3.5	Technical Requirements . . . . .	28
<b>4</b>	<b>Objectives</b>	<b>30</b>
<b>5</b>	<b>Results</b>	<b>31</b>
5.1	LLM comparison . . . . .	31
5.2	Evaluation of Fine-Tuning in Drug Discovery . . . . .	32

5.3	Evaluation of RAG in Drug Discovery . . . . .	34
5.4	Automatic Target Dossier . . . . .	40
5.4.1	PDF . . . . .	40
5.4.2	PowerPoint Presentation . . . . .	43
<b>6</b>	<b>Discussion</b>	<b>51</b>
6.1	Conclusions . . . . .	51
6.2	Perspectives . . . . .	52
	<b>List of Acronyms</b>	<b>55</b>
	<b>List of Annexes</b>	<b>56</b>
	<b>Bibliography</b>	<b>57</b>
	<b>Acknowledgements</b>	<b>73</b>

# List of Figures

2.1	PDAC age-standardised incidence and death rates. From [73].	5
2.2	The Drug Discovery pipeline. From [10].	7
2.3	The Transformer Architecture. From [85].	12
2.4	Evolution process of LM. From [99].	13
3.1	Example of the quantization of a 32-bit floating-point vector into int8 fixed-point numbers. From [51].	19
3.2	RAG: indexing, retrieval and generation. From [28].	21
3.3	Example of fixed-size and semantic chunking.	22
3.4	Illustration of the Naive RAG System.	23
3.5	Illustration of the Advanced RAG System.	24
3.6	Agent architecture. From [89].	25
3.7	Example of the Agent workflow.	27
3.8	Diagram of the LLM Agent system. From [27].	28
5.1	LLM efficiency comparison. From [27].	32
5.2	Box plots summarizing the scores assigned by GPT-4 to the answers of the models. The median score for each metric is represented by the horizontal red line. From [27].	38
5.3	Table of contents of the automatic target dossier PDF. From [27].	41
5.4	Target protein expression in human organs. From [27].	42
5.5	Target role in physiology and tumor progression. From [27].	43

5.6	Target characteristics, including sequence similarity with animals and protein function. From [27]. . . . .	44
5.7	Subcellular location of the target. From [27]. . . . .	45
5.8	Target mutations in cancer patients and general population. From [27]. . . . .	46
5.9	Gene essentiality in human organs. From [27]. . . . .	47
5.10	Existing drugs acting on the target. From [27]. . . . .	48
5.11	SWOT analysis for the target in pancreatic cancer. From [27].	50

# List of Tables

2.1	Hallucinations examples. Text highlighted in red indicates hallucinations, while text highlighted in blue denotes user instructions or provided context that contradicts the LLM hallucination. Table from [41]. . . . .	13
3.1	List of the sources with a short description. Table adapted from [27]. . . . .	17
5.1	Answers of the base and fine-tuned Mistral-7B when asked “What is PDAC?”. . . . .	33
5.2	Answers of the base and fine-tuned Mistral-7B when asked “Is RIPK2 a good therapeutic target for PDAC?”. . . . .	34
5.3	Comparison of the answers of Base and New RAG when asked “Are there any specific biomarkers for diagnosing of PDAC, and if so, which ones?”. . . . .	35
5.4	Comparison of the answers of Base and New RAG when asked “Is RIPK2 a good therapeutic target for PDAC?”. . . . .	37
5.5	Comparison of the answers of the three models when asked “Can you describe the PDAC escape mechanisms?” Table adapted from [27]. . . . .	39

# Chapter 1

## Introduction

Drug discovery is the process of identifying and developing new medications. It consists of identifying, evaluating, and optimizing compounds with activity against a specified target. Drug Discovery includes different procedures such as target identification, target validation and optimization, and finally, the selection of a target for further development [83]. Drug discovery is a very long and expensive process, often taking up to 15 years and requiring approximately 2 billion dollars [8, 36]. The advent of Large Language Models (LLM) brought a paradigm shift to drug discovery, introducing innovative approaches to understanding disease mechanisms, accelerating discovery and enhancing clinical trial optimization [100].

The aim of this thesis is to develop an AI system in order to reduce hallucinations in drug discovery-related tasks. To this end, different adaptation techniques are considered, namely fine-tuning and RAG, highlighting their differences and limitations. After evaluating fine-tuning and RAG, we implemented an LLM-based agent system that generates an automatic target dossier to support the target validation phase of the drug discovery process. A target dossier is an important tool in drug discovery and is used to assess the potential of a gene target for developing a compound that modulates its activity. The target dossier should contain all the relevant information to support the decision-making process of the experts. This includes general information

about the target, its implications in a disease, therapeutic opportunities, and potential competitors. The construction of a target dossier demands considerable time to gather and analyze data about a target from various sources and databases [27]. In this thesis, we demonstrate how it is possible to use an LLM Agent to create an automatic target dossier. Our focus is on pancreatic cancer, which, due to its lethality and malignancy, urgently requires the discovery of new biological targets and therapeutic opportunities.

Our system employs agents in combination with RAG. The agent has different tools at its disposal, allowing it to perform complex tasks and actions such as connecting to external databases, executing Python code, and retrieving images. The system, given an input gene, generates an automatic target dossier in the context of pancreatic cancer, producing a PDF document and a PowerPoint presentation. It retrieves up-to-date information from more than fifteen sources, annotating them in order to allow the verification and cross-checking of the content. The purpose of this system is to support the experts in evaluating a target, a process that requires weeks due to the need to collect all the relevant and reliable information.

Overall, this thesis demonstrates how AI systems can be leveraged to optimize and accelerate the drug discovery process, enabling a quicker development of innovative life-saving therapies.

This thesis is structured as follows:

- **Chapter 2: Background and Related Work.** This chapter explores the complexities and challenges of pancreatic cancer and drug discovery, highlighting the potential of AI to optimize drug discovery tasks. It also covers the rise of Large Language Models, starting with the Transformers architecture and their applications in the medical domain, addressing their limitations and strategies to overcome these challenges.
- **Chapter 3: Materials and Methods.** This chapter introduces the databases, sources, and tools used in this thesis. It provides an overview of the main

techniques employed, specifically Fine-Tuning, RAG, and Agents. Moreover, it lists the key technical requirements essential for the successful implementation of these methods.

- **Chapter 4: Objectives.** This chapter discusses the core objective of the thesis, i.e. the development of an AI system to support the drug discovery process. Furthermore, it details the subgoals required to accomplish it.
- **Chapter 5: Results.** This chapter presents the results of the thesis. The choice of the LLM is discussed, and the methods of Fine-Tuning and RAG are evaluated. Finally, the target dossier produced by the system is presented, showing examples of the Agent process to generate some PDF pages and PowerPoint slides.
- **Chapter 6: Discussion.** This chapter summarizes the key findings of the thesis and offers insights into how the Agent system could be improved through the integration of additional tools or the adoption of multimodal machine learning.

# Chapter 2

## Background and Related Work

### 2.1 AI need for Drug Discovery

#### 2.1.1 Pancreatic Ductal Adenocarcinoma

Pancreatic ductal adenocarcinoma (PDAC), the most frequent subset of pancreatic cancer, is a highly lethal gastrointestinal tumor. It occurs when abnormal DNA mutations in the pancreatic ductal cells lead to uncontrolled growth and division, resulting in the formation of tumors [52]. PDAC is a leading cause of cancer-related mortality, with a poor overall prognosis that has shown no significant improvement for many decades [55]. According to the National Cancer Institute, in 2024, pancreatic cancer was the 10th most common cancer based on estimated new cases in the United States, but ranked 3rd as the deadliest cancer, with over 50,000 fatalities. The deadliness of PDAC is also reflected by the 5-year survival rate, which fluctuates between 12% and 13%. Figure 2.1 shows the age-standardised incidence and death rates of pancreatic cancer. It can be observed that most of the people diagnosed with the disease do not survive.

Its prevention and early diagnosis are challenging. Indeed, the majority of patients with pancreatic cancer remain asymptomatic until the disease has progressed to an advanced stage, and there is no established standard screening

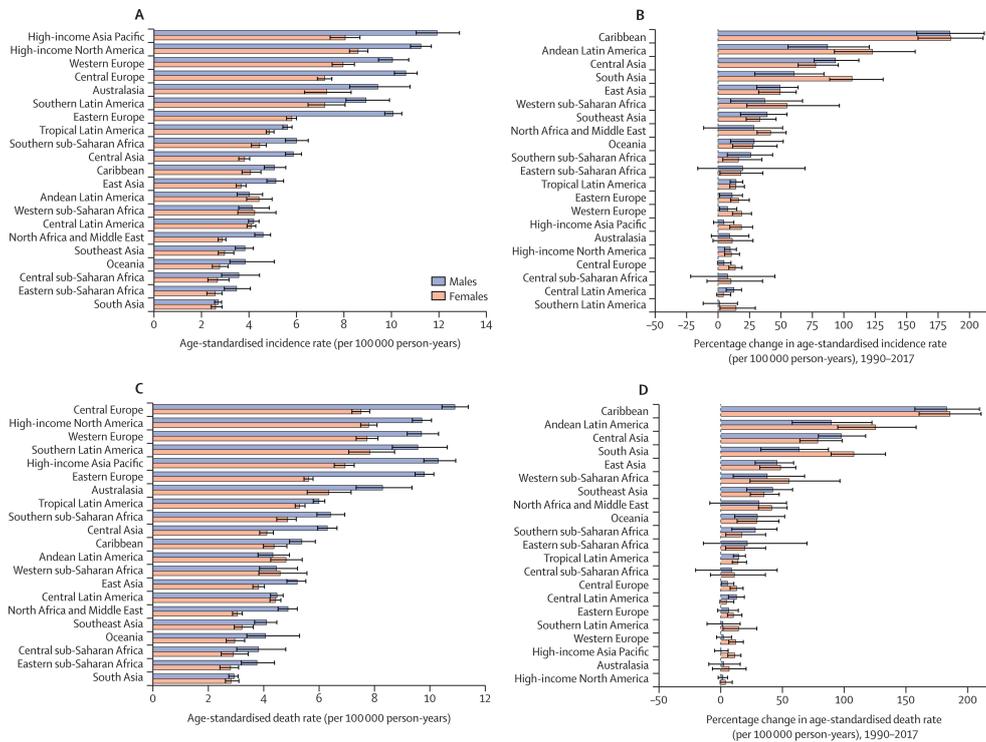


Figure 2.1: PDAC age-standardised incidence and death rates. From [73].

program for individuals at high risk of PDAC [53]. To this day, surgical resection of the tumor is regarded as the only treatment option with the potential to cure pancreatic cancer [80]. Over the past decade, two new combination regimens have become the standard first-line treatment in patients with advanced pancreatic cancer. The first one, known as FOLFIRINOX, is a combination of 5-fluorouracil (5-FU), leucovorin, irinotecan, and oxaliplatin. The second regimen combines gemcitabine with an albumin nanoparticle conjugate of paclitaxel [33]. In a clinical trial [17] it was discovered that patients treated with FOLFIRINOX have a higher overall median survival compared to patients treated with gemcitabine. However, the overall median survival remains very low (around 11%), highlighting the need for new therapeutic solutions.

The key driver genes in pancreatic cancers include KRAS, CDKN2A, TP53, and SMAD4/DPC4, with KRAS mutations found in more than 90% of PDAC patients. These genes undergo mutations at various stages of precursor lesions, and their dysregulation facilitates the differentiation and proliferation

of pancreatic cancer cells [39]. However, some of these genes are hardly drug-gable, while others are commonly resistant to inhibitors [22]. For instance, KRAS, the most commonly mutated gene in pancreatic cancer, is considered "undruggable" [94], a term that refers to target proteins with flat functional interfaces that lack defined pockets for ligand interaction, presenting a significant challenge for rational drug design. KRAS is considered undruggable because therapeutic molecules cannot bind effectively to its small binding pockets, combined with a highly competitive GTP concentration, which makes the development of a KRAS inhibitor challenging [30]. PDAC therapeutic options are limited, and advancements in drug development are hindered due to the genomic, epigenetic, and metabolic complexities of most pancreatic cancers [55]. For these reasons, it is crucial to identify new therapeutic targets, a process that is achieved through drug discovery.

### 2.1.2 The Drug Discovery Process

Drug discovery is the process of identifying and developing new medications. It consists of identifying, evaluating, and optimizing compounds with activity against a specified target. The process [82, 86] typically starts by focusing on a disease and identifying potential targets, generally gene products, that can be affected by small compounds. These compounds are intended to either interfere with or prevent the disease or, at the very least, slow the progression of symptoms. There are many techniques to identify therapeutic targets, including cellular assays, genomic studies, and proteomic studies. The initial identification of candidate compounds typically relies on high-throughput screening of a wide variety of small molecule collections or structurally selected compounds that are either known to have activity against a target or are predicted to do so. Promising compounds are tested on the basis of many criteria, including specificity, toxicity, pharmacology, and biopharmaceutical properties. Then, some of them are evaluated in animal models or in vitro

models. Meanwhile, studies on absorption, distribution, and elimination are conducted. After several years of research, a handful of compounds may be considered safe and effective enough to advance to patient trials to assess the safety and efficacy of the drugs in patients. The studies are then submitted to regulatory agencies, which evaluate the documents and determine whether the compound should be approved for the market. If the review is favorable, it is possible to release the drug to the market and prescribe it to patients. After approval, monitoring is conducted to track any potential side effects that may arise over time from the new treatment. The Drug Discovery process is illustrated in Figure 2.2.



Figure 2.2: The Drug Discovery pipeline. From [10].

The development of a new drug must progress through multiple stages to ensure it is safe, effective, and meets all regulatory requirements [19]. Drug discovery is a very long and expensive process, requiring up to 15 years and approximately 2 billion dollars [8, 36]. Moreover, most of the compounds fail the tests. On average, for every 5,000 to 10,000 tested compounds, only one is approved [25]. Compounds can fail for a variety of reasons. Among them, toxicity is a leading cause of failure, accounting for 30% of all failures. Another key cause of failure is a lack of efficacy, which also has a percentage

rate equal to 30% [69].

### 2.1.3 AI in Drug Discovery

Recently, Artificial Intelligence techniques have gained significant attention in the drug discovery pipeline due to their potential to accelerate the process by efficiently analyzing large amounts of data [11]. As Deep Learning (DL) technology continues to advance and drug-related data expands, approaches based on deep learning are increasingly being applied across all stages of drug development [4]. The potential of AI in drug discovery has been extensively demonstrated through various studies. In [103], a deep learning-based efficacy prediction system (DLEPS) was developed to identify drug candidates by using changes in gene expression profiles associated with the diseased state. DLEPS was proven to be capable of providing valuable insights into pathogenic mechanisms. Another application of AI in drug discovery is demonstrated in [23], where DL was used to extract informative features from genome-scale omics data and to train classifiers to predict the effectiveness of drugs in cancer cell lines.

Due to the high failure rate in the drug discovery process, particularly because of target toxicity and low efficacy, tools to reduce this risk are needed. Collecting information about the target can be very helpful as it allows for an analysis of the relevance of the target in a specific disease, as well as the identification of possible toxicity-related risks. This may save time and resources needed. A tool that collects all the relevant information about a target is the target dossier. The target dossier is an important instrument in drug discovery as it is used to assess the suitability of a target that has already been identified. The target dossier should contain all the relevant information to support the decision-making process of the experts. This includes general information

about the target, its role in a disease, therapeutic potential, and potential competitors. The construction of a target dossier requires significant time to collect and analyze data about a target from different sources and databases [27]. Automating the target dossier using AI would save the experts a significant amount of time and would therefore facilitate the process of drug discovery. Despite the increasing number of AI applications in the drug discovery domain, to our knowledge there is currently no system that creates an automatic target dossier.

## **2.2 Large Language Models in Medicine**

### **2.2.1 Transformers**

Large Language Models could help in the automation of the target dossier since they are able to analyze, process, and generate textual data.

LLMs are built upon the transformer architecture, introduced in 2017 in the paper “Attention Is All You Need” [85]. Transformers leverage several key features that contribute to their remarkable effectiveness. In Natural Language Processing (NLP) tasks, raw text cannot be processed by a model. Thus, it undergoes the process of tokenization, which converts human-readable text into a sequence of distinct tokens (word subdivision) [79]. Tokens can be generated at different levels of granularity, such as characters, subwords, or entire words. Typically, models that use character-level tokenization tend to perform worse as capturing semantic relationships becomes more challenging, making tokenization at the subword level a more common choice [76]. Moreover, subword-level tokenization solves the issue of out-of-vocabulary (OOV) words. This term refers to words that are not seen at training time by the model, which is therefore unable to deal with them. Historically, rare words that were not present in the vocabulary were replaced by the UNK (unknown) token. However, this token is unacceptable in tasks like natural

language generation [64]. With subword-level tokenization, the OOV rate is zero [65], since rare words can be obtained by combining known subwords. Subwords are then encoded into token ids [76], as models are unable to operate on text and can only process numerical data. Tokens are mapped into high-dimensional vectors called embeddings, which encode syntactic and semantic information, capturing the tokens' meaning and relationships inside a sentence [3]. Before passing the token embeddings to the Transformer model, additional information is required: the positions of the tokens in the input sequence. This is because, unlike recurrent neural networks (RNNs) or Long Short-Term Memory networks (LSTMs), Transformers do not process the input sequentially; they are permutation equivariant [14]. This information is provided by positional encodings. The token embeddings, enriched with positional information, are then passed as input to the Transformer model. At the core of the Transformer architecture is the concept of Attention (derived from the human equivalent, it is meant to assign higher weight to more important words in a context). The self-attention mechanism determines the level of attention assigned to other words when encoding the word at the current position [34]. This enables the model to capture long-range dependencies and represent the interactions among all the tokens in the sequence [2]. The Transformer architecture employs a Multi-Head Self-Attention mechanism composed of multiple parallel self-attention heads, allowing the model to capture and learn different relationships within the sequence. The Transformer architecture typically has an encoder-decoder structure. It is illustrated in Figure 2.3. The scope of the encoder is to analyze the contextual information of the input, while the decoder generates the output by using the encoder's output and masked multi-head attention [44]. The encoder consists of multiple blocks, each containing two fundamental layers: the Multi-Head Attention Layer and Feed-Forward Layer. The latter layer comprises two linear transformations and a non-linear rectified linear unit (ReLU) activation function, which is applied independently to each position. This enables the model to

learn the complex transformations of the data at each position [77]. In addition, the encoder block integrates residual connections around both layers, together with layer normalization [44]. The purpose of residual connections is to address the vanishing gradient problem by allowing the gradient to flow more easily through the network [81]. In addition, layer normalization normalizes activations of intermediate layers, enabling smoother gradients, faster training, and better generalization accuracy [95]. The decoder has a structure similar to the encoder, but it adds a third sub-layer that applies Multi-Head Attention over the output of the encoder. Moreover, the first Multi-Head Attention layer is masked. This masking restricts positions from attending to subsequent positions in the sequence [85].

### 2.2.2 Large Language Models

LLMs are based on the Transformer architecture but go beyond it by substantially increasing both the model size and the amount of data that are used for training. For instance, they can contain hundreds of billions of parameters and are trained on a vast collection of textual data [99]. The evolution of language models, from statistical language models to LLMs, is illustrated in Figure 2.4.

LLMs exhibit enhanced language understanding and generation capabilities, along with new emergent abilities [66]:

- In-context learning, which allows LLMs to learn new tasks given only a few examples in the prompt as demonstrations [24],
- Instruction following, which refers to the ability of LLMs to follow natural language instructions, also known as zero-shot prompts [102], i.e. to perform a task without the need for examples,
- Multi-step reasoning, the LLMs' capability to generate intermediate reasoning steps in order to solve complex problems, often referred to as chain-of-thought [88].

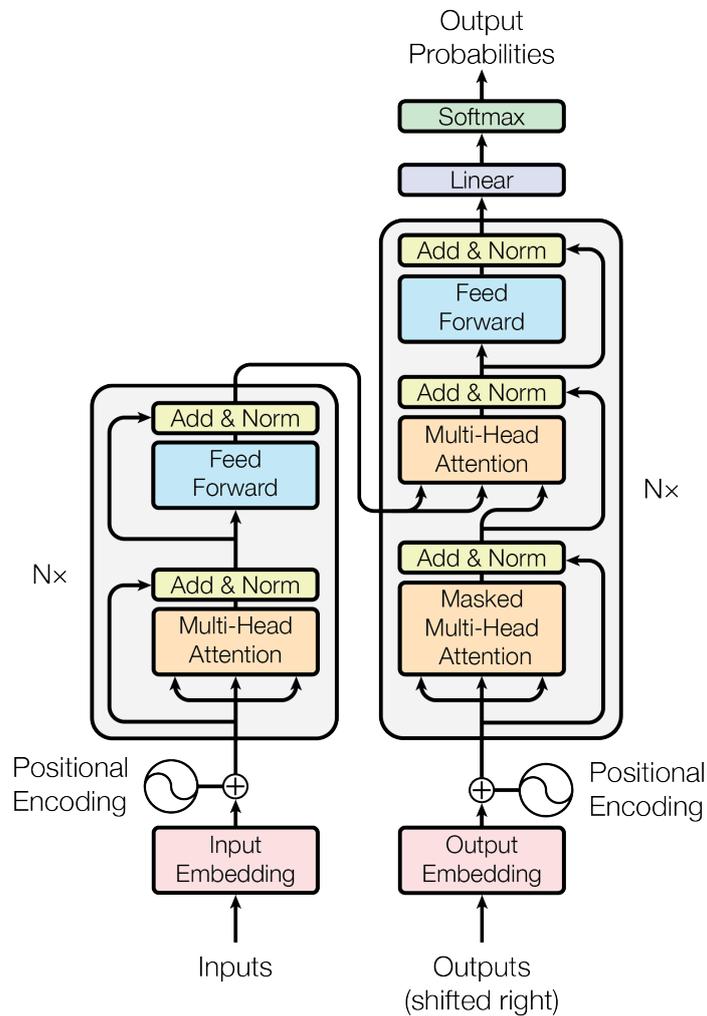


Figure 2.3: The Transformer Architecture. From [85].

Despite their remarkable capabilities, LLMs are subject to some limitations that need to be addressed in order to fully leverage their potential. One of the most important challenges is the issue of LLM hallucinations which result in seemingly plausible yet factually incorrect content that appears highly convincing and human-like [41]. This is mainly observed in domain-specific applications since LLMs are trained on a vast range of general data and may lack subject-specific knowledge. Hallucinations in LLMs typically occur when the model tries to fill in knowledge or context gaps, by making assumptions based on patterns it has learned during training. This can result in incorrect or misleading outputs, which can be problematic in sensitive applications [32]. Some

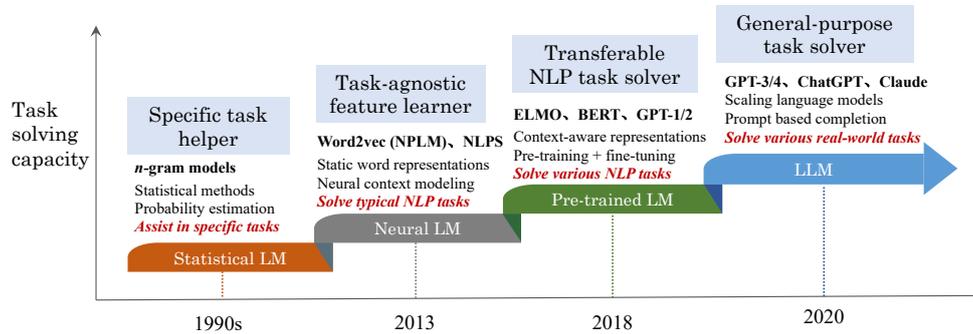


Figure 2.4: Evolution process of LM. From [99].

examples of hallucinations in LLMs are illustrated in Table 2.1.

Category	Type	Example Dialog	Explanation
Factuality Contradiction	Factual	<b>Instruction:</b> What are Thomas Edison's main contributions to science and technology? <b>Response:</b> Thomas Edison developed the first practical telephone and invented the light bulb.	The response is factually incorrect. In reality, Edison improved the light bulb, building on earlier designs by others, and Alexander Graham Bell invented the telephone.
	Factual Fabrication	<b>Instruction:</b> What are the major environmental impacts of the construction of the Eiffel Tower? <b>Response:</b> The construction of the Eiffel Tower in 1889 led to the extinction of the Parisian tiger, a species that played a crucial role in the region's ecosystem. Additionally, it is widely recognized as the event that sparked the global green architecture movement.	The response erroneously states that the construction of the Eiffel Tower led to the extinction of the 'Parisian tiger' and sparked the global green architecture movement. In fact, there is no evidence of a 'Parisian tiger' ever existing, making this a fabricated claim. Moreover, attributing the origins of green architecture to the Eiffel Tower is an exaggeration, as this movement has diverse roots and cannot be traced back to a single event.
Faithfulness Hallucination	Instruction Inconsistency	<b>Instruction:</b> Translate the English question into Spanish: "What is the capital of France?" <b>Response:</b> The capital of France is Paris.	The user explicitly provided an instruction to translate the English question into Spanish. Instead of performing the translation as instructed, the LLM chose to answer the question.
	Context Inconsistency	<b>Retrieved context:</b> The Nile is a major north-flowing river in northeastern Africa. It's traditionally considered the longest river in the world. Originating in the Great Lakes region of central Africa, it flows through several countries before emptying into the Mediterranean Sea. <b>Instruction:</b> Summarize the retrieval document above. <b>Response:</b> The Nile originates from the mountain ranges of central Africa and flows into the Mediterranean.	The summary incorrectly states that the Nile originates from the "mountain ranges of central Africa", while the paragraph mentions that the Nile originates in the Great Lakes region of central Africa.
	Logical Inconsistency	<b>Instruction:</b> Solve the equation $2x + 3 = 11$ step by step. <b>Response:</b> Step 1: Subtract 3 from both sides to get: $2x = 8$ . Step 2: Divide both sides by 2 to get: $x = 3$ . So the result of the equation is $x=3$ .	The first step correctly isolates the term with x to get $2x = 8$ . However, the next step inaccurately divides 8 by 2 to yield a result of $x = 3$ , which is inconsistent with the earlier reasoning.

Table 2.1: Hallucinations examples. Text highlighted in red indicates hallucinations, while text highlighted in blue denotes user instructions or provided context that contradicts the LLM hallucination. Table from [41].

Another major issue is that the knowledge of LLMs is limited to the data used for training them, meaning they lack awareness of events and information that surfaced after the training period. For instance, they are unable to access real-time data.

### 2.2.3 The Application of LLMs in Drug Discovery

The advent of LLMs such as GPT [75] and BERT [21] has increased interest in leveraging LLMs in the biomedical field, including applications in drug discovery [12, 100]. However, the above issues must be addressed with caution, particularly in sensitive fields such as healthcare, where an error may have severe consequences. To mitigate these risks, significant efforts have been made to adapt LLMs for the biomedical domain, in order to improve accuracy in medical question-answering and minimize hallucinations. In [15], the authors refined a series of LLMs based on the Llama-2 architecture to enhance medical knowledge retrieval, reasoning, and question-answering abilities. The results show that refined models outperform their corresponding base models across all medical benchmark datasets. To enhance the reliability of LLMs in specialized domains, a widely adopted technique is Fine-Tuning. This process involves taking a pre-trained LLM and further training it on a smaller, domain-specific dataset. This approach transfers the pre-trained model's learned patterns and features to new tasks, improving performance on specific tasks while reducing data and computational requirements [71]. Fine-Tuning helps reduce hallucinations. However, like pre-trained LLMs, a fine-tuned model lacks real-time knowledge. Keeping the model updated would require frequent fine-tuning, which is impractical, particularly for LLMs with tens of billions of parameters, due to the high computational costs. Moreover, fine-tuning may lead to catastrophic forgetting, a phenomenon that refers to the LLM's tendency to lose previously acquired knowledge when learning new data [40]. In addition, if the target domain for fine-tuning is too distant from the original pre-training domain, the fine-tuning process can lead to a substantial increase in hallucinations. An alternative approach to reduce hallucinations is RAG [28], a solution that incorporates knowledge from external databases. RAG consists of retrieving relevant documents from external

knowledge databases through semantic similarity and providing these documents as context for the LLM to enhance the accuracy of the answer. The documents contain relevant information for answering the user query or question and help in reducing the possibility of producing incorrect answers. RAG is more efficient than fine-tuning, since in order to integrate new information it is only necessary to update the knowledge database, avoiding model training [7]. However, like fine-tuning, RAG lacks real-time knowledge, as it requires updating the database. A promising approach could be to combine RAG and LLM Agents. Agents employ tools that allow them to perform more complex actions, such as accessing online resources and querying different databases via API [89], as well as processing and plotting data. The target dossier is a crucial step in the drug discovery process, so it should contain reliable and accurate information. Moreover, it must remain up-to-date with the latest advancements and discoveries, tracking newly published biomedical literature and database updates. A RAG and Agent-based approach could be very effective and beneficial in this application.

# Chapter 3

## Materials and Methods

### 3.1 Data

As presented in Section 2.1, the target dossier is an important instrument in drug discovery for assessing the suitability of a target in a specific disease. To create it, several pieces of information to support the decision-making process of the experts are needed. To gain a complete view of the target and allow the LLM to engage in high-level reasoning, it is necessary to collect data from various sources. One of the most important sources in biomedical sciences is PubMed Central (PMC) [26], a free full-text archive of biomedical and life sciences literature. PMC is used in this work as a data source to support the RAG process, enabling the LLM to generate paragraphs about specific characteristics of a target. Another crucial data source is UniProt [18], a comprehensive resource of protein sequences that contains information such as the target function, subcellular location, and sequence. Fundamental details can also be found in The Human Protein Atlas [5], an open-access resource containing protein data such as RNA and cell-line expression in both healthy and cancerous tissues. All the data sources and tools used in this work are listed in Table 3.1.

<b>Source</b>	<b>Description</b>
UniProt	A database offering high-quality and freely accessible resource for protein sequence.
Human Protein Atlas	A database focused on genome-wide analysis of human proteins.
DrugBank	An online database containing information on drugs and drug targets.
Open Targets	A database that uses human genetics and genomics data for drug target identification and prioritization.
RCSB PDB	An online database containing information on proteins.
cBioPortal	A resource for the exploration of cancer genomics datasets containing information such as mutation frequencies.
TCGA Survival	A website that provides analysis of mutations, copy number alterations, etc., associated with cancer outcome in TCGA.
OGEE	An online database containing information about gene essentiality.
STRING	A database about protein-protein interactions including both direct and indirect associations.
SIGNOR	A repository of annotated causal relationships among human proteins, chemicals of biological relevance, stimuli and phenotypes.
ESMO	The European Society for Medical Oncology provides oncology information, including guidelines on cancer.
PubChem	A database that contains information about molecules such as chemical structures, chemical and physical properties and biological activities.
Gene	An NCBI database that integrates gene information from different species.
PubMed	An NCBI database comprising millions of citations and abstracts about biomedical literature.
PMC	An NCBI archive of biomedical and life sciences journal literature.
SNP	An NCBI databases containing information about single nucleotide mutations.
BLAST	A tool that finds the similarity between biological sequences.
DeepTMHMM	A Deep Learning Model for classification and prediction of Transmembrane Topology.
GSEApY	A Python package to perform Gene Set Enrichment Analysis.

Table 3.1: List of the sources with a short description. Table adapted from [27].

The sources are also specified in the target dossier PDF and PowerPoint generated by the agent. This allows the reader to trace back the sources that have been used to obtain information about a specific section or slide. The references are specified at the end of each section in the PDF and in the speaker notes of each PowerPoint slide. For example, the paragraphs generated with RAG include notes specifying the IDs of the PMC articles used as context for generating them. When the information is collected from online databases, a direct link to the information is added (when available).

## 3.2 Model Fine-Tuning

Fine-tuning is the process of taking a pre-trained model and further training it on a smaller and domain-specific dataset [71]. LLMs are trained on extensive datasets with unsupervised learning objectives to learn general language representations. However, their performance can be suboptimal when applied to specific tasks, such as question answering, as they lack necessary domain knowledge. Fine-tuning offers an effective solution to improve their performance in these tasks [96]. However, LLMs are trained on huge datasets and learn billions of parameters. Performing full fine-tuning, which updates the weights of all layers, would require extensive time and resources. Instead, Parameter-Efficient Fine-Tuning (PEFT) consists of freezing some of the layers of the pre-trained model and fine-tuning only some additional parameters, which are tailored to the specific task [45]. The PEFT method used in this work is Low Rank Adaptation (LoRA). LoRA enables the training of some dense layers in a neural network indirectly. Instead of directly optimizing all the weights of dense layers, LoRA adjusts their rank decomposition matrices, which represent changes in the dense layers during fine-tuning [38]. Additionally, these LoRA low-rank matrices can be combined to enable cross-task generalization, supporting multi-task learning, domain adaptation, and continual learning [63]. Moreover, big models are usually quantized to reduce

the computational and memory requirements. Quantization consists of transforming the model’s weights and activations from higher-precision formats, such as 16-bit floating-point numbers, to lower-precision formats, such as 8-bit or 4-bit, to preserve the model’s performance while substantially reducing its memory footprint [56]. Figure 3.1 illustrates an example of quantization, where a 32-bit floating-point vector is converted into int8 fixed-point numbers. The values are mapped to the integer range  $[-128, 127]$ , preserving relative differences in magnitude.

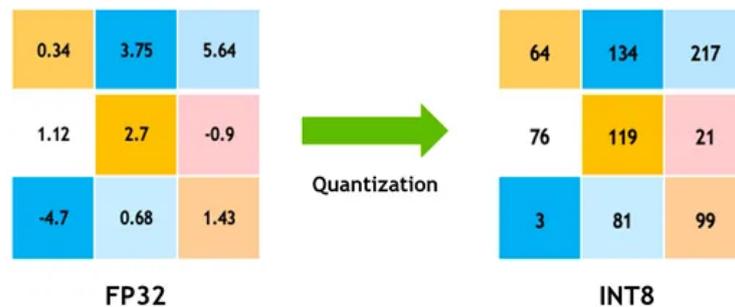


Figure 3.1: Example of the quantization of a 32-bit floating-point vector into int8 fixed-point numbers. From [51].

In this work, the LLM (Mistral-7B) was quantized to 8-bit and fine-tuned using LoRA. The scope was to enhance Mistral-7B’s knowledge for answering pancreatic cancer-related questions. The fine-tuning dataset was composed of abstracts of biomedical publications about pancreatic cancer coming from PMC. Each observation in the dataset consisted of a pair of instructions and an answer. We generated different possible tasks to train the model (in agreement with the instruction-tuning approach). The aim was to create tasks that respect the biological domain and the instruction-tuning format (question followed by the answer). There were different possible instruction-answer options. The instruction could involve summarizing the abstract of an article, and in this case, the answer would be its title (the rationale is that the title of the article could be seen as a summary of its abstract). Another option for the instruction could be

to extract the keywords from an abstract. The answer would be the keywords of the article that can be found in the article's metadata. Moreover, in order to prevent overfitting to our data, 20% of the fine-tuning dataset consisted of a random sample from *Gath\_baize* [29], a dataset that contains instructions and answers about a wide range of topics. In addition, adding general knowledge prevents the model from catastrophic forgetting and too narrow adaptation to the domain knowledge. In total, the fine-tuning dataset contained 12,000 observations. The fine-tuning process of Mistral-7B required approximately eight hours.

However, as anticipated in Section 2.2, fine-tuning remains computationally expensive and is impractical in applications that require frequent information updates. Moreover, it may lead to catastrophic forgetting.

### 3.3 Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is an innovative approach designed to overcome some of LLMs' limitations. It integrates two key components: a retrieval mechanism, which retrieves relevant documents from an external knowledge base, and a generative model (the LLM), which processes this retrieved information to produce a reliable and coherent response [31]. The RAG system consists of three main steps: indexing, retrieval, and generation [28]. These three phases are clearly illustrated in Figure 3.2.

Indexing starts with the extraction of data, which is possibly converted into textual data. Due to the limited context of LLMs (the maximum number of tokens they can handle), text is divided into chunks. There are different chunking techniques [98, 101]. The fixed-size strategy splits text into equal segments but often neglects the inherent structure of the text. On the other hand, the recursive strategy divides the text incrementally, using separators like punctuation marks, to better adapt to the content. Another approach is the contextual strategy, which segments the text based on its meaning and

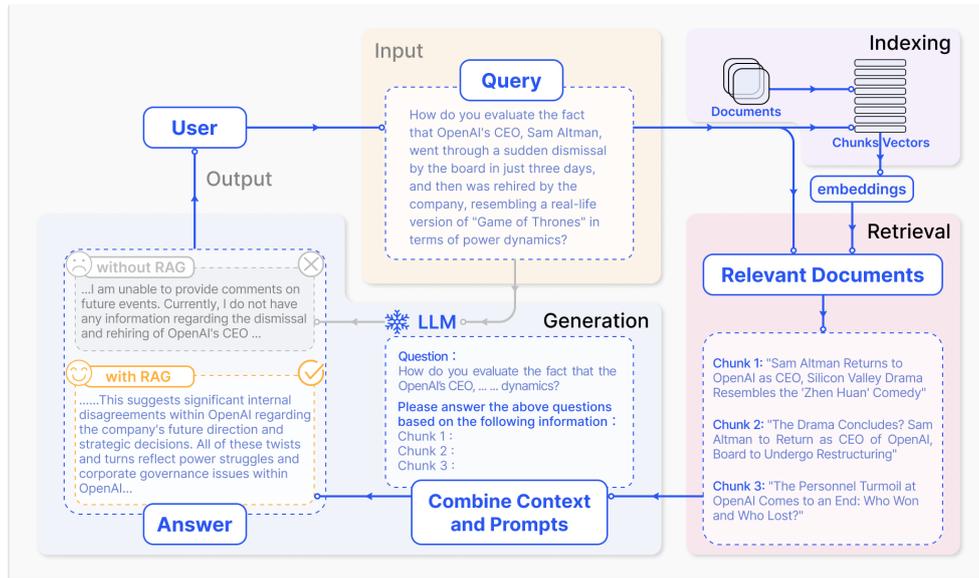


Figure 3.2: RAG: indexing, retrieval and generation. From [28].

structure to capture the meaning in context (semantic chunking). There are also more advanced techniques, such as Parent Document Retrieval [58], that use small chunks for retrieval while returning larger blocks for the generation phase. This is because embeddings of smaller texts are more effective at capturing their semantic meaning, while longer texts help retain contextual meaning and long dependencies during generation. An example of fixed-size and semantic chunking is illustrated in Figure 3.3. The text in the example is extracted from [68].

Fixed-size chunking divides the text into equal portions, truncating words and overlapping sentences. On the other hand, semantic chunking divides the text into structurally and semantically coherent segments. For this reason, the latter technique is preferred and implemented in this work.

Chunks are then transformed into feature vectors using an embedding model and stored in a vector database. During the retrieval phase, the user query is transformed into a vector using the same embedding model from the previous step. The system computes semantic similarity scores between the query vector and the stored embeddings in the vector database. The top  $K$  chunks

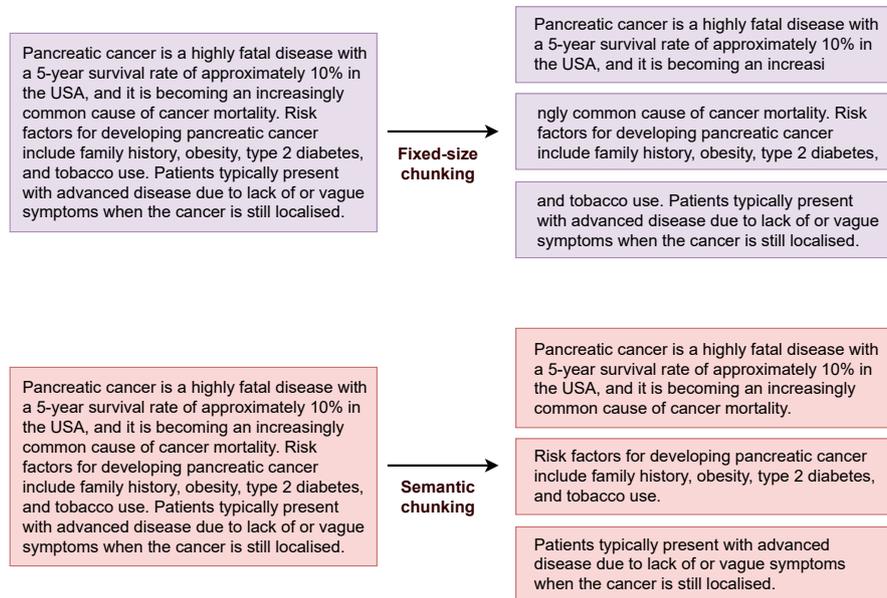


Figure 3.3: Example of fixed-size and semantic chunking.

with the highest similarity score are retrieved and passed to the LLM as context, which then produces a response (generation phase). This approach is known as Naive RAG. The first simple RAG system that we implemented is illustrated in Figure 3.4.

This naive RAG system is very simple, relying only on semantic search and an embedding model. In medical applications, there are many specific terms to refer to different pathologies, therapies, and biological concepts. For this reason, during the RAG search phase, instead of considering only semantic similarity, it could be worth performing a similarity search based on both semantic meaning and keywords (hybrid search). An information retrieval algorithm that implements keyword search is BM25, using the term frequency-inverse document frequency (TF-IDF). It computes the similarity score of a chunk based on how frequently a query term appears in that chunk, accounting for the document's length and the frequency of the term in the corpus

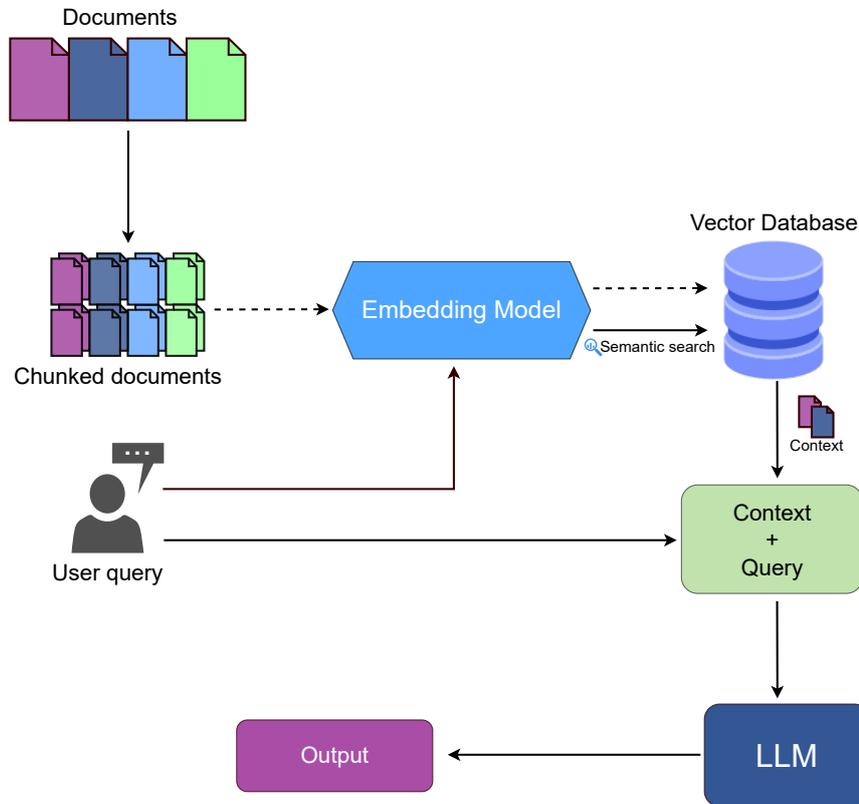


Figure 3.4: Illustration of the Naive RAG System.

[31]. Moreover, in the Naive RAG system, the embedding model must compute the similarity score between the query and each chunk. Due to the usual high number of chunks, the embedding model should be relatively small for efficiency reasons, and its limited capacity could result in a non-optimal retrieval process. An alternative is to integrate a Reranking model [74] in the system. In this new approach, the embedding model retrieves a higher number of chunks which are then reranked by the reranking model. The reranker is usually a pre-trained transformer that has been fine-tuned for this specific task. This allows the reranking model to be bigger and, therefore, more performant than the embedding model: it only needs to process the chunks that have already been retrieved. The top  $K$  most similar chunks are selected by

the reranker and passed as context to the LLM. This new advanced RAG system that integrates a hybrid search and a reranking step is illustrated in Figure 3.5.

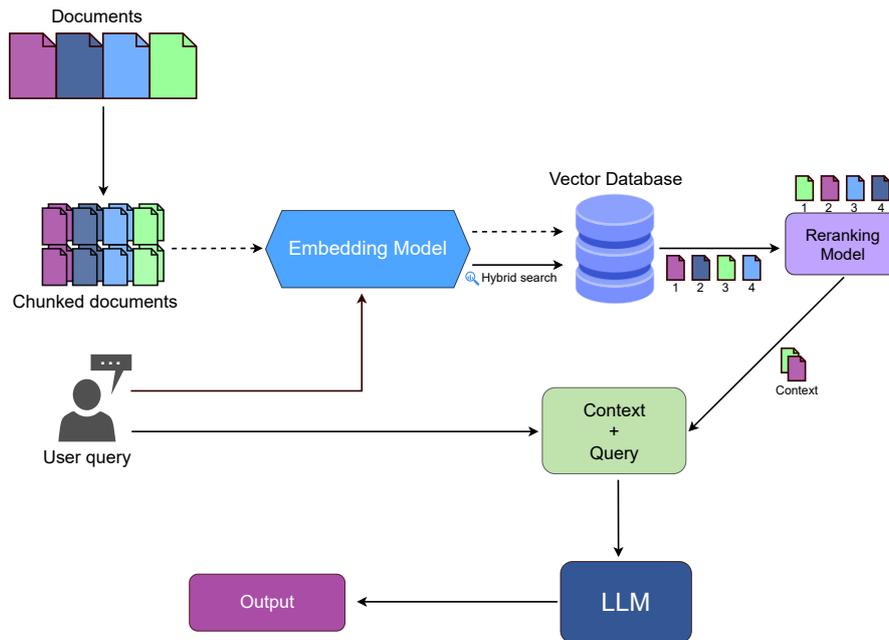


Figure 3.5: Illustration of the Advanced RAG System.

In this work, we used LangChain Semantic Chunker [57] in order to split documents based on semantic similarity. The chunks were then embedded using *bge-base-en* [42], a relatively small open-source embedding model, and they were stored in a temporary collection using Chroma [16] Vector Database. For the reranking process, *mxbai-rerank-large* [43] was employed.

### 3.4 LLM Agents

Large Language Models have shown emergent reasoning abilities, enabling them to perform multi-step reasoning to generate answers. However, this “chain-of-thought” reasoning is a static black box, relying solely on the internal representations of the model to generate thoughts. Since it is not grounded

in the external world, its ability to reason reactively or incorporate new knowledge is limited [97]. To overcome these limitations, it is possible to employ LLMs as the brain part of the system and integrate them with additional components. The agent architecture is illustrated in Figure 3.6.

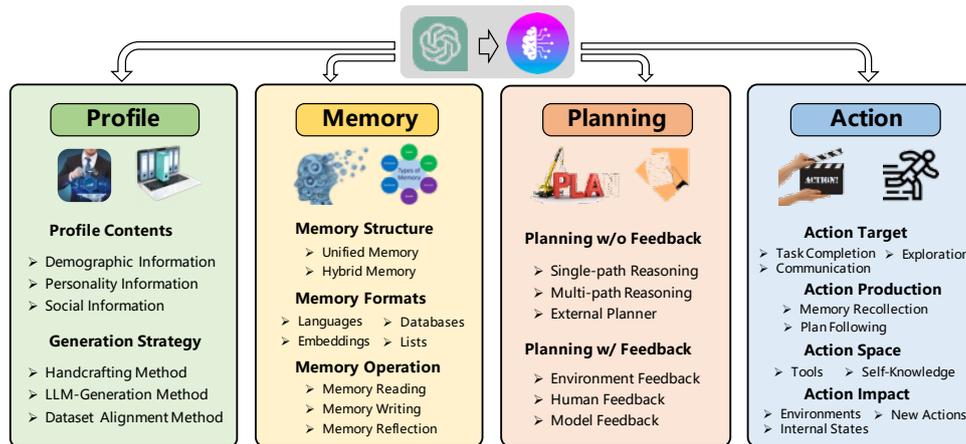


Figure 3.6: Agent architecture. From [89].

The essential components of an agent are [91]:

- **Planning:** At the center of planning, there is reasoning. Through reasoning, agents can decompose a complex problem into multiple manageable sub-tasks, developing suitable plans for each one. As tasks progress, agents can use introspection (the ability of the agent to evaluate its past actions) to adjust their plans, making sure they align better with real-world circumstances and improving the final output [93].
- **Memory:** The memory module allows the agent to store information perceived from the environment and use the stored memories to facilitate and guide future actions. The memory module can help the agent to collect experiences, self-evolve, and act in a more consistent, rational, and effective way [89]. Memory mechanisms allow the agent to recall and implement previous strategies effectively when dealing with complex problems. Moreover, these memory mechanisms allow the agent to adapt to new environments by recalling past experiences [93].

- **Tools utilization:** The utilization of external tools allows the agent to expand its action space and perform tasks with higher quality [93]. The Agent can invoke other models or use tools to handle specific sub-tasks, particularly in areas where LLMs face challenges, such as calculations and real-time data retrieval [90]. Tools allow the agent to access databases through APIs, run Python code, search for information on the internet, and perform multiple other tasks. The prompt of the LLM includes the list of available tools, along with guidelines on when to use each one.

Our Agent utilizes a prompt template based on ReAct [97] to incorporate explicit steps to guide the LLM's thinking process. It includes three steps:

- **Thought:** the agent reasons about what to do.
- **Action:** the tool that the agent decides to use.
- **Observation:** the information retrieved using the tool.

Once the Agent arrives at the final answer to the user's query, it returns the response, and the process ends. Figure 3.7 illustrates an example of the Agent workflow.

Due to their ability to perform complex actions and retrieve accurate information, LLM Agents can be employed for the creation of an automatic target dossier. In this work, we implemented a system that, given a target gene and a type of cancer, creates a target dossier as a PDF and a PowerPoint presentation with all the required information to assess the effectiveness of the target in that specific cancer. In order to create the target dossier, the Agent has different tools at its disposal that allow it to:

- Produce texts and summaries about the target and cancer characteristics using the advanced RAG system presented in Section 3.3,
- Plot protein and molecule structures, as well as retrieve relevant images from different databases,

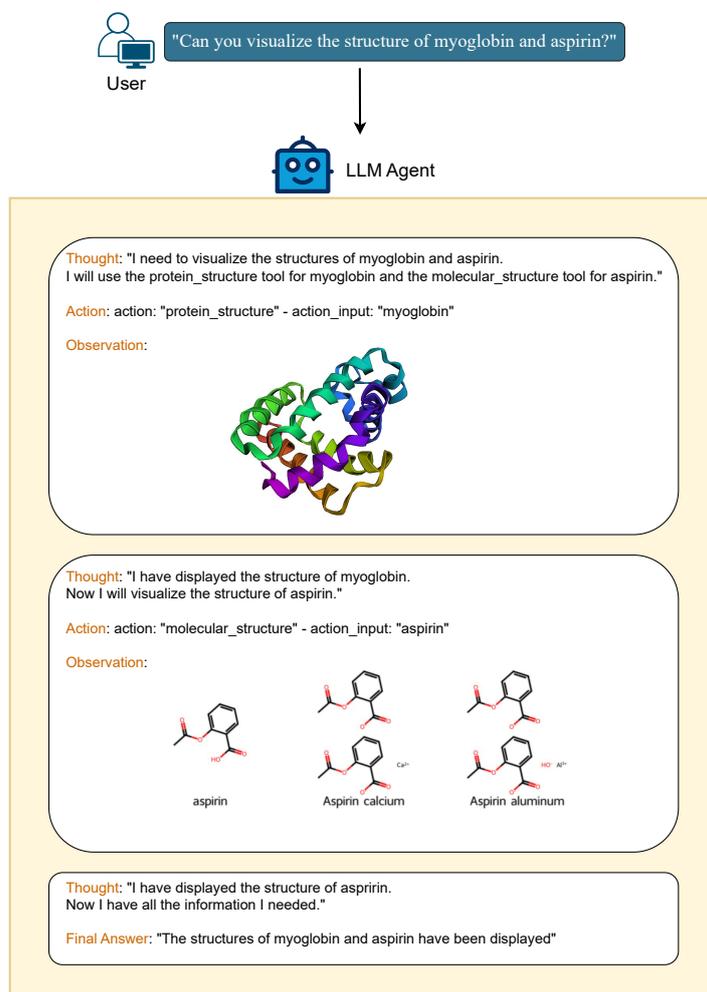


Figure 3.7: Example of the Agent workflow.

- Run Python code to connect with APIs, retrieve and process data, and produce plots and tables,
- Perform a final analysis in order to determine the suitability of a target given the retrieved information.

Once the Agent has collected all the necessary information, the target dossier PDF is generated, followed by a PowerPoint presentation, which is a summarized version of the PDF. As stated in Section 3.1, each section and slide

contains notes on the original data sources. The system is summarized in Figure 3.8.

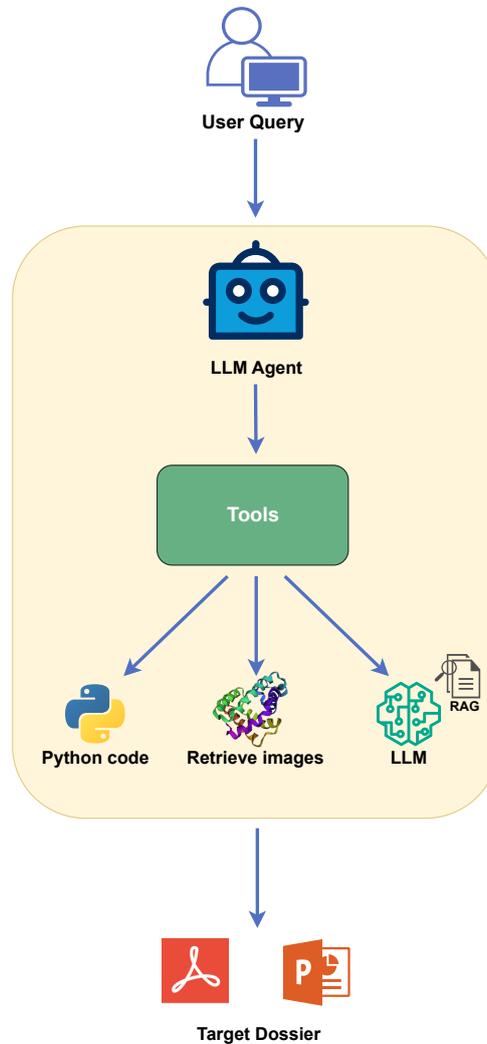


Figure 3.8: Diagram of the LLM Agent system. From [27].

### 3.5 Technical Requirements

The main platforms and applications employed in this work are HuggingFace [92], LangChain [13], and the Chroma database[16]. HuggingFace is a platform that provides the *Transformers* library, focused on supporting Transformer-based architectures and simplifying the distribution of open-source pre-trained

models [92]. HuggingFace was used for loading and deploying all the models, including LLMs, embedding models, and rerankers. LangChain is a framework for developing applications powered by LLMs. LangChain played a central role in this work since it provides many libraries that were used for different crucial implementations. Indeed, LangChain libraries were used to create document chunks, support the RAG process, and develop the LLM Agent together with its tools. Moreover, it enables an easy integration with HuggingFace. Finally, Chroma was chosen as the vector database to store the document embeddings and perform retrieval in the RAG process. Chroma is easily integrable with LangChain, making it an excellent choice. This work was developed in Jupyter Notebook, using version 3.11.5 of Python. The fine-tuning and model deployments were carried out using eight Tesla V100 GPUs, each with 32 gigabytes of memory.

# Chapter 4

## Objectives

The main objective of this thesis is to develop an AI system to support and accelerate the drug discovery and development process, using methods such as Agents and RAG to reduce hallucinations and improve the quality and performance of the system. For these reasons, we focused on a deadly cancer as PDAC, which is in dire need of new therapeutic options. Moreover, we focused on the target dossier where the use of RAG and LLM agents can be beneficial. To achieve this, multiple subgoals need to be defined and addressed. They include:

- Selecting an open-source LLM to serve as brain component of the agent, focusing the evaluation on the quality of the generated output and generation speed.
- Implementing and evaluating techniques to enhance the LLM's knowledge of drug discovery and reduce hallucinations, specifically fine-tuning and RAG, through examples.
- Implementing an Agent system that incorporates the use of external tools and RAG to generate an automatic target dossier, with the aim of supporting the target validation step in the drug discovery process.

# Chapter 5

## Results

### 5.1 LLM comparison

The HuggingFace *Transformers* library supports the distribution and utilization of various pre-trained models in a centralized model hub [92]. It contains many open-source models that can be deployed and adapted to specific tasks. Besides the quality of the answers generated by LLMs, a crucial aspect is their efficiency, the speed at which tokens are generated. When creating an automatic target dossier, the LLM needs to generate long texts, which can result in thousands of tokens. For this reason, the LLM must be relatively fast. We focused our model choice on open-source models with a permissive license, and with a middle range for parameters. Several LLMs were compared, including Mistral-7B-instruct [46], Qwen-1.5-7B-Chat [6], Orca 13B [67], Gemma-7B-instruct [84], Mixtral-8x7B-instruct [47]. These models range in size from 7 billion to 13 billion parameters, except Mixtral-8x7B-instruct, which has 46.7 billion parameters. The LLMs were evaluated on a predefined set of pancreatic cancer-related questions, considering both the quality of their responses and their efficiency in token generation. All the models produced good answers, with Mixtral-8x7B-instruct demonstrating exceptional performance in this task. However, it was very slow, generating an average of 0.73 tokens per second on our GPU setup. To create a 500-token paragraph, Mixtral would

require approximately 11-12 minutes, leading to an excessively high computational time when generating multiple sections. An optimal balance between effectiveness and efficiency was achieved by Mistral-7B-instruct [46], which was approximately 13 times faster than Mixtral. For this reason, Mistral-7B-Instruct was chosen as the base LLM for this work. Figure 5.1 shows the average tokens generation per second of all the LLMs under evaluation.

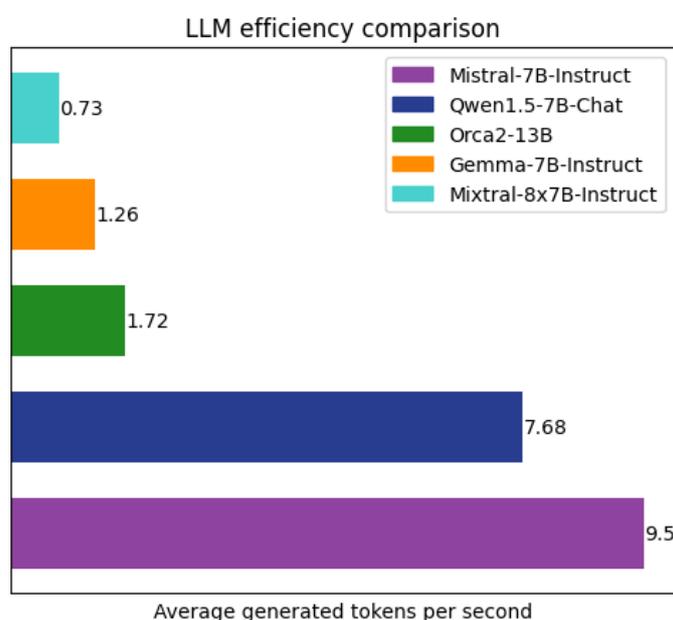


Figure 5.1: LLM efficiency comparison. From [27].

## 5.2 Evaluation of Fine-Tuning in Drug Discovery

Mistral-7B was fine-tuned using the setup described in Section 3.2 to enhance its ability to answer questions related to pancreatic cancer. The base Mistral-7B and its fine-tuned version were evaluated and compared on pancreatic cancer-related questions. Table 5.1 shows the answers of the models to the question “What is PDAC?”, which is the acronym of Pancreatic Ductal Adenocarcinoma.

The response generated by Mistral-7B is entirely incorrect as it fails to

<b>What is PDAC?</b>	
<b>Mistral-7B</b>	PDAC stands for the Professional Development and Accreditation Council. It is a national accrediting agency in Canada that evaluates and accredits post-secondary education programs in business and technology. The agency’s primary goal is to ensure that the education provided in these programs meets the needs of the industry and prepares students for success in their chosen fields...
<b>Fine-Tuned Mistral-7B</b>	PDAC stands for Pancreatic Ductal Adenocarcinoma, which is a type of cancer that develops in the cells of the pancreas. It is the most common type of pancreatic cancer and is characterized by rapid growth and spread to other parts of the body. Symptoms may include abdominal pain, weight loss, and jaundice. Treatment options include surgery, chemotherapy, and radiation therapy.

Table 5.1: Answers of the base and fine-tuned Mistral-7B when asked “What is PDAC?”.

recognize that pancreatic cancer is the subject of the question. Instead, it mistakenly interprets PDAC as referring to the Professional Development and Accreditation Council, which is unrelated to the context of the question. In contrast, the fine-tuned model provides a correct yet brief answer. Table 5.2 shows another example. When prompted with the question “Is RIPK2 a good therapeutic target for PDAC?”, base Mistral-7B’s answer is not very specific and lacks important information. In particular, it mentions that further research is needed to assess the safety and effectiveness of RIPK2 without addressing potential challenges. In contrast, the fine-tuned model generates a brief introduction about RIPK2, explains its implications in pancreatic cancer, and, most importantly, addresses the challenges in targeting it, delivering a comprehensive overview of its potential and limitations.

Despite the good results achieved through fine-tuning, the process remains computationally expensive and impractical in a domain like drug discovery, which requires up-to-date information.

<b>Is RIPK2 a good therapeutic target for PDAC?</b>	
<b>Mistral-7B</b>	RIPK2 is a promising therapeutic target for pancreatic adenocarcinoma (PDAC). It is involved in the regulation of cell death and inflammation, and has been shown to be overexpressed in PDAC. Inhibition of RIPK2 has been shown to reduce tumor growth and improve survival in preclinical models of PDAC. However, further research is needed to determine the safety and efficacy of RIPK2 inhibitors in clinical trials.
<b>Fine-Tuned Mistral-7B</b>	RIPK2 is a key regulator of the innate immune response and has been implicated in the pathogenesis of pancreatic adenocarcinoma (PDAC). RIPK2 is overexpressed in PDAC and is associated with poor prognosis. Therefore, targeting RIPK2 may be a promising therapeutic strategy for PDAC. However, there are several challenges to consider when targeting RIPK2 in PDAC. First, RIPK2 is a key regulator of the innate immune response and has a role in the development of cancer. Therefore, targeting RIPK2 may have unintended consequences on the immune system. Second, RIPK2 is also involved in other pathways beyond the innate immune response, such as cell death and autophagy. Therefore, targeting RIPK2 may have off-target effects on other cellular processes. In conclusion, while RIPK2 is a promising therapeutic target for PDAC, further research is needed to determine the safety and efficacy of targeting this protein in this disease.

Table 5.2: Answers of the base and fine-tuned Mistral-7B when asked “Is RIPK2 a good therapeutic target for PDAC?”.

### 5.3 Evaluation of RAG in Drug Discovery

In Section 3.3, two RAG models were presented: Naive RAG and Advanced RAG, which integrates a hybrid search and a reranker. In this case as well, the models’ responses to the same set of questions were compared. Table 5.3 shows the answers to an example question. Base (Naive) RAG is unable to answer the question because the documents retrieved by the system

<b>Are there any specific biomarkers for diagnosing of PDAC, and if so, which ones?</b>	
<b>Base RAG</b>	<b>The context provides information about the current clinical standard for diagnosing pancreatic ductal adenocarcinoma (PDAC) and the need for less invasive biomarkers that can be used in combination with current strategies. However, it <b>does not provide a specific answer to the question about which biomarkers are used for diagnosing PDAC.</b></b>
<b>New RAG</b>	<b>The context provides information about the current FDA-approved biomarker for PDAC diagnosis, which is Carbohydrate antigen 19-9 (CA19-9).</b> However, the context also mentions that CA19-9 has limitations, such as variable specificity and the inability to express CA19-9 at all in some individuals. The context also mentions that traditional imaging methods have limited value for early diagnosis of PDAC .Therefore, the context suggests that there is a need for more effective biomarkers with higher sensitivity and specificity for early stage PDAC. The context also mentions that metabolomics is a promising tool for discovering valuable diagnostic biomarkers and understanding carcinogenesis. <b>Several studies have demonstrated the potential use of metabolites as diagnostic biomarkers for PDAC.</b> <b>The context also provides information about a specific biomarker panel that has been identified as potentially useful for diagnosing PDAC. This panel includes proline, creatine, and palmitic acid.</b> The panel showed significant diagnostic performance in the training and validation sets, with AUC values of 0.854 and 0.865, respectively, in the comparison of PDAC and Ctr (control). In summary, the context suggests that there are specific biomarkers for diagnosing PDAC, including proline, creatine, and palmitic acid, but more research is needed to validate these biomarkers and develop a reliable diagnostic panel for early stage PDAC.

Table 5.3: Comparison of the answers of Base and New RAG when asked “Are there any specific biomarkers for diagnosing of PDAC, and if so, which ones?”.

through semantic search do not contain any relevant information. This is due to the utilization of only a small embedding model, which is not very effective. Moreover, the lack of keyword search may be a limitation, as the question contains specific terms like *biomarkers* and *PDAC*, which could help in document retrieval. This limitation is overcome by Advanced (New) RAG, which is capable of identifying different biomarkers, such as Carbohydrate antigen 19-9 (CA 19-9), which, as stated in [72], is a commonly used biomarker for the diagnosis of pancreatic cancer. Another example is shown in Table 5.4. Also, in this case, Base RAG cannot answer the question based on the retrieved documents. This again highlights the importance of a reranking model and keyword search, especially in the biomedical field, which contains many technical terms (like *RIPK2* in this query). Advanced RAG is able to find much information about the potential of *RIPK2* as a target for pancreatic cancer, as supported by several studies [78, 48].

To assess the effectiveness of our RAG system, three different models were compared: Mistral-7B without RAG, Mistral-7B with Naive (Base) RAG, and Mistral-7B with Advanced (New) RAG. However, assessing the answers generated by LLMs is not trivial. The evaluation was performed by GPT-4, which was specifically prompted to assign a score to each question based on five different metrics: faithfulness, relevance, quality, completeness, and correctness [27]. The results of the evaluation are summarized in the box plots shown in Figure 5.2. The results show that Mistral-7B with New RAG outperforms the other models. The median score is equal to 5 out of 5 for each metric, and 75% of the average scores are equal to or greater than 4, with the lowest achieved score being 3. On the other hand, the median score of the other two models is equal to 4, and they achieve a score of 1 out of 5 on some questions. Overall, Mistral-7B with Base RAG outperforms Mistral-7B without RAG, as 75% of its average scores are equal to or above 4, compared to the model without RAG, which has a value of the first quartile (Q1) equal to 3. Table 5.5 shows the answers of the three models to a sample question. Similarly

<b>Is RIPK2 a good therapeutic target for PDAC?</b>	
<b>Base RAG</b>	<p><b>The provided context does not mention RIPK2 as a potential therapeutic target for PDAC.</b></p> <p>It discusses CDK1 expression as a potential stratification factor for patient therapy, and the use of CDK7 inhibition in combination with standard chemotherapy as a possible novel treatment regimen for overcoming PDAC chemoresistance.</p>
<b>New RAG</b>	<p><b>Based on the provided context, RIPK2 has been found to function as a crucial pro-tumor gene in pancreatic cancer (PC) and can become a potential intervention target in PC treatment.</b></p> <p>The dysregulated RIPK2 expression in PC affects the expression of other members of RIPKs, especially RIPK1 and RIPK3, which can induce necroptosis to drive pancreatic cancer progression. RIPK2 can also activate the autophagy and suppress reactive oxygen species (ROS) production, which is intimately linked to the kinase activity of RIPK2 itself. RIPK2 self-assembled endosome or RIPosome provides a signal of being eaten in the process of autophagy, which may serve as a scavenger to eliminate ROS in cells. Increased ROS production can trigger pancreatic cancer cell apoptosis. RIPK2 knockout can suppress the formation of autophagosomes, elevate ROS level and promote PC cell apoptosis. RIPK2 interacts with PRKCI, which also has a higher expression in PC tissues than the normal, and a higher PRKCI expression predicts a poorer survival of PC patients. PRKCI activates NF-<math>\kappa</math>B and JNK signaling to promote PC cell growth, migration and invasion. RIPK2 enhances the immune score, stromal score and ESTIMATE score in PC. RIPK2 promotes a series of immune cell infiltration, especially in KIRC, TGCT and THCA. RIPK2 significantly enhanced the immune score, stromal score and ESTIMATE score in PAAD, PCPG and KICH. These findings suggest that RIPK2 is closely related to tumor immune infiltration.</p>

Table 5.4: Comparison of the answers of Base and New RAG when asked “Is RIPK2 a good therapeutic target for PDAC?”.

to the question in Table 5.1, Mistral-7B is unable to recognize PDAC as the subject of the question. Instead, it discusses the Pacific Data and Automation

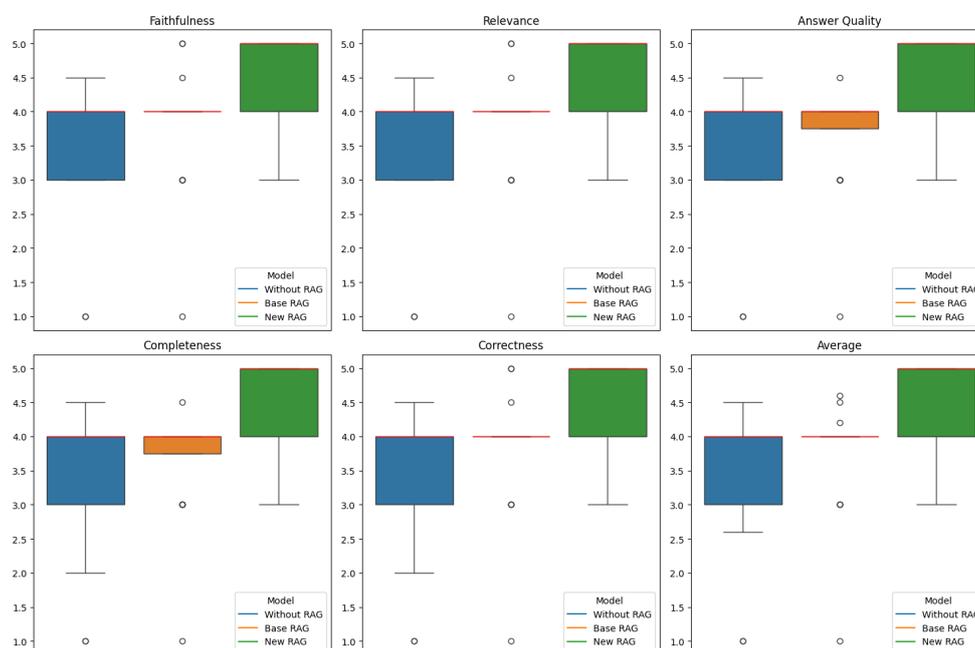


Figure 5.2: Box plots summarizing the scores assigned by GPT-4 to the answers of the models. The median score for each metric is represented by the horizontal red line. From [27].

Corporation escape mechanisms, mistakenly interpreting the acronym PDAC in relation to our application. The answers of the two RAG models are similar, as both identify two main escape mechanisms: the use of Gasdermin E (GSDME) [62] to form a barrier that prevents chymotrypsin-mediated destruction, and the sequestration of Major Histocompatibility Complex class I (MHC-I) [9] inside PDAC cells through autophagy. Additionally, the Advanced RAG system identifies a third evasion factor, tumor-derived HLA-G [37].

<b>Can you describe the PDAC escape mechanisms?</b>	
<b>Without RAG</b>	<p><b>The PDAC (Pacific Data and Automation Corporation)</b> escape mechanisms are a set of security features designed to prevent unauthorized access to the data stored on a PDAC device.</p> <p>These mechanisms include:...</p>
<b>Base RAG</b>	<p><b>The provided context discusses two possible escape mechanisms for Pancreatic ductal adenocarcinoma (PDAC):</b></p> <ol style="list-style-type: none"> <li>1. Gasdermin E (GSDME) and mucin expression: PDAC cells use GSDME to form a barrier that prevents chymotrypsin-mediated destruction. GSDME facilitates the expression of mucin 1 and mucin 13, which form this barrier. Inhibition or knockout of mucin 1 or mucin 13 abrogates orthotopic PDAC growth in NOD-SCID mice.</li> <li>2. Autophagy and Major Histocompatibility Complex class I (MHC-I) sequestration: Autophagy was recently described as a tumor-intrinsic immune escape process during tumor development by sequestering MHC-I inside PDAC cells. This discovery implies that autophagy may be a potential target for limiting immune escape in patients and improving immunotherapy efficiency.</li> </ol> <p>It is important to note that the context does not provide enough information to fully understand the PDAC escape mechanisms or their implications for treatment. Further research and studies are needed to fully understand and address these mechanisms.</p>
<b>New RAG</b>	<p><b>Based on the provided documents, it appears that pancreatic ductal adenocarcinoma (PDAC) has several escape mechanisms</b> that allow it to evade the immune system and survive in the pancreatic microenvironment. One such mechanism involves the use of gasdermin E (GSDME) to form a barrier that prevents chymotrypsin-mediated destruction. GSDME facilitates the expression of mucin 1 and mucin 13, which form this barrier. Inhibition or knockout of mucin 1 or mucin 13 abrogates orthotopic PDAC growth in NOD-SCID mice. Another mechanism involves the sequestration of Major Histocompatibility Complex class I (MHC-I) inside PDAC cells through autophagy. This process supports tumor progression and may be a potential therapeutic target. Additionally, tumor-derived HLA-G may indicate the mechanism of immune escape and impaired PDAC clinical outcome, especially in early-stage patients. Overall, understanding these escape mechanisms is crucial for the development of more effective immunotherapies for PDAC.</p>

Table 5.5: Comparison of the answers of the three models when asked “Can you describe the PDAC escape mechanisms?” Table adapted from [27].

## 5.4 Automatic Target Dossier

The LLM Agent, which generates an automatic target dossier as described in Section 3.4 and Figure 3.8, was tested using different genes in the context of pancreatic cancer. To demonstrate the system's capabilities, the automatic target dossier for KRAS (KRAS was chosen for this test as it is a well-known target in PDAC) is presented. The target dossier is generated as two different files: a PDF document and a PowerPoint presentation. Both files contain the utilized data sources and the date on which the dossier was produced. Some PDF pages and PowerPoint slides are shown below. The complete PDF and PowerPoint presentation are available here: <https://github.com/Oncodesign-Precision-Medicine/Automatic-Target-Dossier>.

### 5.4.1 PDF

Figure 5.3 shows the table of contents of the target dossier PDF. It is divided into four different sections. The first section provides various information about the target such as its expression, the mutations present in cancer patients and healthy individuals, its essentiality, interactions, and role in pancreatic cancer tumor progression. The second section is focused on the disease of interest, pancreatic cancer, in this case. It presents different information which includes a description of the disease, some statistics, and the ESMO guidelines to help decide the treatment for a patient with cancer. The third section is about existing cancer therapies or drugs that act on the target of interest. Finally, the last section contains a SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis and a conclusion to evaluate whether the target could be a good candidate for pancreatic cancer treatment. Figure 5.4 shows the target protein expression under wild-type conditions, i.e., the form in which it is found in nature. The data on expression in different human organs are retrieved by the Agent by connecting to the Human Protein Atlas database. After the extraction, the data are plotted in Python to visualize expression as



### Table of contents:

1. Target information.....	3
1.1 Summary and characteristics.....	3
1.2 Transmembrane Helix Prediction.....	5
1.3 Subcellular location.....	6
1.4 Expression.....	7
1.5 Mutations.....	11
1.6 Glycosylations.....	15
1.7 Gene essentiality.....	16
1.8 Protein-protein interactions.....	19
1.9 Pathway enrichment.....	20
1.10 SiGnaling Network.....	21
1.11 Role in physiology.....	25
1.12 Role in tumor progression.....	25
1.13 Kaplan-Meier curves.....	26
2. Disease information.....	34
2.1 Disease description.....	34
2.2 Disease statistics.....	35
2.3 ESMO guidelines.....	37
3. Competitive landscape.....	43
3.1 Pancreatic cancer standard of care.....	43
3.2 Pancreatic cancer current therapies.....	44
3.3 Known drugs targeting KRAS.....	45
4. Conclusion.....	53
4.1 SWOT analysis.....	53
4.2 Conclusion.....	55

Figure 5.3: Table of contents of the automatic target dossier PDF. From [27].

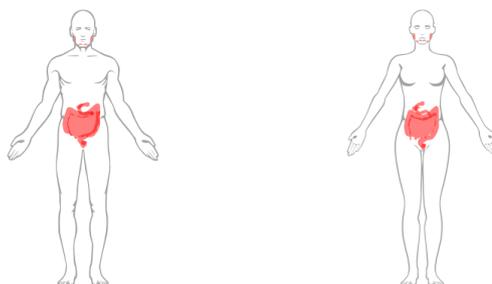
colored human body images. In particular, the first image highlights in red the human organs in which the target is highly expressed in both male and female individuals. On the other hand, the second image shows the expression levels in each organ. The darker the red highlighting an organ, the higher the target expression level. It is possible to notice that KRAS is highly expressed in the digestive system, suggesting its relevance in pancreatic cancer. The page contains indications about the meaning of the colors to help the reader interpret the images. However, the data source is not visible, as it is provided on

the following page at the end of the target expression subsection. Figure 5.5

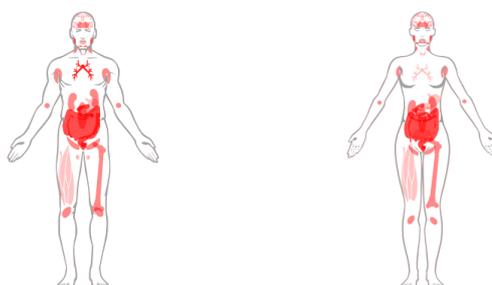
#### 1.4 Expression

##### Wild type KRAS protein expression levels

The following image shows organs in which KRAS protein expression level is high.



The following image shows KRAS protein expression level in the organs. The darker the colour, the higher the expression level.



KRAS is highly expressed in Appendix, Colon, Duodenum, Gallbladder, Rectum, Salivary gland, Small intestine

Figure 5.4: Target protein expression in human organs. From [27].

contains information about the role of the target in physiology and tumor progression, presented in two different paragraphs. The texts are generated using the advanced RAG system explained in Section 3.3, with one RAG for each subsection. The contexts gathered during the two retrieval processes consist of chunks from PMC articles. The generated paragraphs provide comprehensive and well-explained descriptions to help experts gain insights into the role of

the target in physiology and tumor progression. In case of doubt or to double-check the information, the source used is indicated below the paragraphs. It points to the references page, which contains a list with the IDs of the PMC articles that were used as context for generation.

#### 1.11 Role in physiology

KRAS is a GTPase that plays a fundamental role in transducing signals from plasma membrane growth factor receptors to downstream signaling pathways controlling cell proliferation, survival, and migration. In normal cells, KRAS activity is tightly controlled, but specific mutations, such as those at codons 12, 13, or 61, disrupt the RAS protein's ability to transition between its active and inactive states, leading to persistent activation and uncontrolled cell growth, contributing to various cancers. KRAS signaling also plays an essential role in regulating the balance of secretory, ciliated, and squamous cell differentiation of the human airway epithelium. In the context of the provided study, siRNA-mediated knockdown of KRAS decreased differentiation of basal stem/progenitor cells into secretory and ciliated cells, while activation of KRAS signaling via lentivirus-mediated over-expression of the constitutively active G12V KRAS mutant had the opposite effect. Cigarette smoke exposure increases KRAS and RAS protein family activation in vitro and in vivo, contributing to airway epithelial remodeling.

Source: PubMed [10]

#### 1.12 Role in tumor progression

KRAS is a frequently mutated oncogene in human adenocarcinoma of the lung and pancreas. In pancreatic cancer, KRAS mutations are linked with poor clinical outcomes. Recent studies suggest that KRAS-driven tumor progression may not be completely independent of upstream signaling, as ERBB family receptor tyrosine kinases (RTKs) have been shown to play a role in KRAS-driven lung tumor development and progression. Similarly, in pancreatic ductal adenocarcinomas (PDAC), SMAD4 deficiency, which is associated with KRAS mutations, has been shown to accelerate PDAC development and alter tumor phenotype. Additionally, a non-receptor protein tyrosine phosphatase, SHP2, has been identified as an essential player in oncogenic KRAS-driven tumors, and its inhibition or deletion has been shown to delay tumor progression but not achieve tumor regression. Synergy is observed when both SHP2 and MEK are targeted, resulting in sustained tumor growth control in murine and human patient-derived organoids and xenograft models of pancreatic ductal adenocarcinoma. These findings suggest that KRAS plays a central role in pancreatic cancer progression and that targeting multiple signaling pathways may be an effective therapeutic approach for KRAS-mutant cancers.

Source: PubMed [11]

Figure 5.5: Target role in physiology and tumor progression. From [27].

### 5.4.2 PowerPoint Presentation

After the generation of the PDF, a PowerPoint presentation is automatically created. The presentation can be seen as a summarized version of the PDF,

designed to present the information clearly and concisely. Moreover, the presentation is easily editable, allowing for the correction of possible incorrect or incomplete information. Figure 5.6 illustrates the slide containing some characteristics of the target, namely the similarity between the target human protein sequence and that of some animals, as well as the function of the protein. The sequence similarity is very important because it provides indications of which organisms a drug could be tested on, i.e., organisms that have a protein sequence similar to the human one. The creation of this slide involves different processes to extract and organize the information, involving the use of APIs and external tools. First, the agent connects to the UniProt API and searches for information about the target protein sequence. A separate search is performed for each organism's sequence. After the extraction of the protein sequences, the similarity between the human and animal sequences is calculated using BLAST (Basic Local Alignment Search Tool). The output of BLAST, along with the target protein function extracted from UniProt, is then organized into a table for better visualization. The sources are listed at the bottom left of the slide. The speaker notes contain a link pointing to the UniProt page of the target. Figure 5.7 provides insights into the subcellular location of

#### Target characteristics

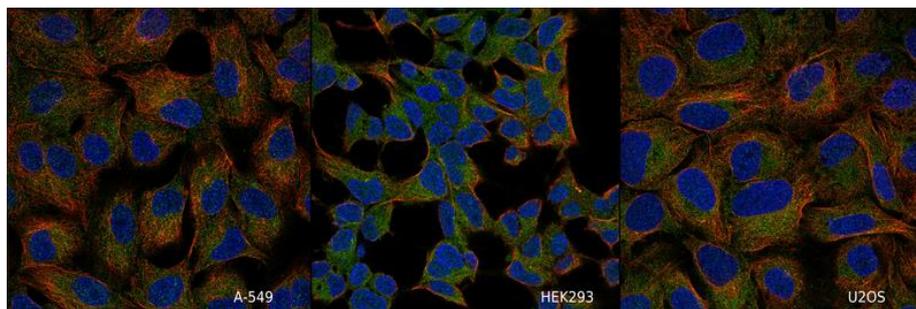
Target characteristics	
Similarity with monkeys	97.01%
Similarity with mice	98.94%
Similarity with rabbits	98.94%
Similarity with dogs	98.94%
Similarity with Guinea pigs	100.0%
Protein function	Ras proteins regulate cell proliferation by binding GDP/GTP and inducing TSG silencing in CRC cells through ZNF304.

Source: UniProt and BLAST

Figure 5.6: Target characteristics, including sequence similarity with animals and protein function. From [27].

the target. The Agent extracts the images of cell lines from the Human Protein Atlas, adds the name of the corresponding cell line to the bottom right corner of each picture, and concatenates the images horizontally. A legend indicating the meaning of the colors is added. The target is represented in green, while blue and red represent the nucleus and microtubules, respectively. Moreover, the main location of the target (cytosol in the case of KRAS) and the antigen used for the analysis are specified. The speaker notes on the slide contain a direct link to the web page from which the data was retrieved.

#### Subcellular location



- Localized to the cytosol.
- Antigen: HPA072761

- KRAS
- Nucleus
- Microtubules

Source: Human Protein Atlas

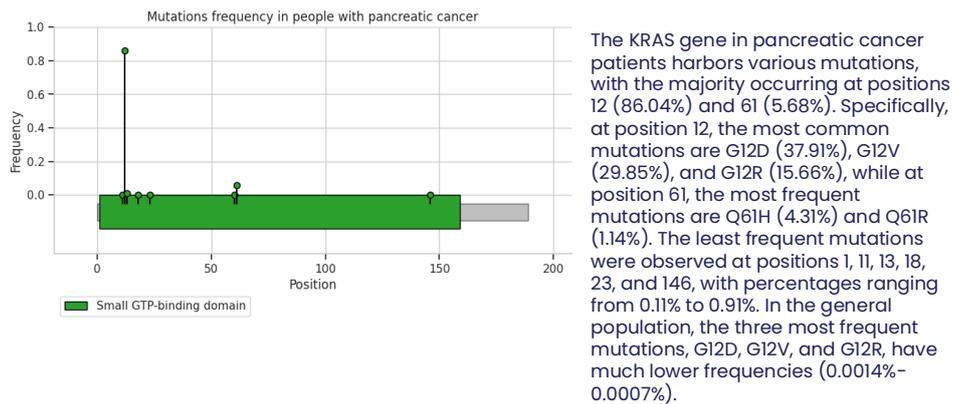
opm

Figure 5.7: Subcellular location of the target. From [27].

Figure 5.8 provides details about the target mutations in patients with the cancer of interest and the general population. This information is crucial because if the mutation rate in cancer patients is much higher than in healthy individuals, the target is very likely correlated with the disease. To collect data on mutations, the agent connects to cBioPortal and retrieves the frequency and location of mutations in cancer patients. Data on mutations in the general population are obtained from the National Center for Biotechnology Information (NCBI) SNP database. The mutation frequencies and locations are organized into a table in the PDF. However, tables can be very long and need to be summarized in the PowerPoint presentation. This is achieved by the LLM, which

is specifically prompted to generate a summary of the mutations table, which is converted into text. Besides the summary, a lollipop chart is plotted in Python in order to provide a clear visual representation of the frequency and location of mutations in cancer patients. The plot also includes the target domains. The findings in this slide support the involvement of KRAS in pancreatic cancer. Indeed, the KRAS mutation frequency is very high, with mutations at location 12 occurring in more than 86% of cancer patients. In contrast, mutations at this position in the general population are very rare, with a percentage frequency ranging between 0.0007% and 0.0014%.

#### Mutations frequency in pancreatic cancer



Source: cBioPortal



Figure 5.8: Target mutations in cancer patients and general population. From [27].

Another important aspect when evaluating a target is its essentiality (Figure 5.9). Essential genes are those that are necessary for the survival of an organism [49]. To collect this information, the Agent accesses the OGEE database, which contains results from gene essentiality experiments. It uses Python to generate a stacked bar chart representing the percentage of tested cell lines in which the target is essential, specifying the reference dataset. Moreover, it generates another bar chart to illustrate gene essentiality in each organ, adding a small paragraph that lists the organs in which the target is essential

(using a threshold of 50%).

### Gene essentiality

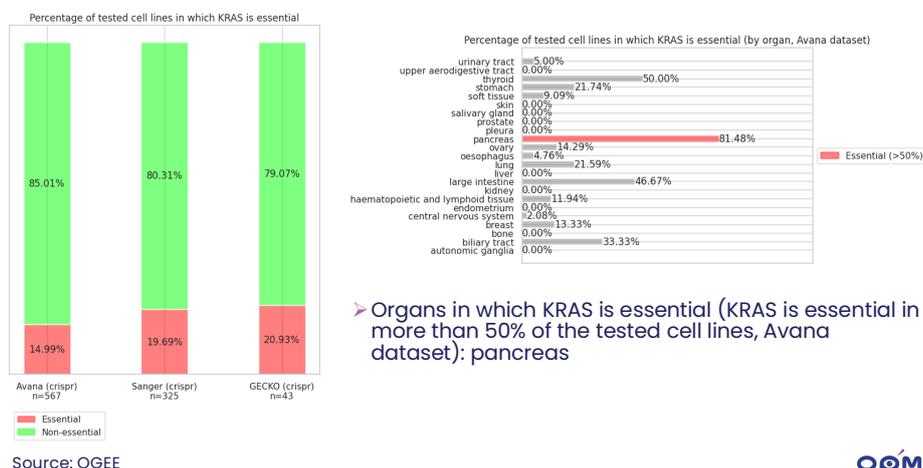
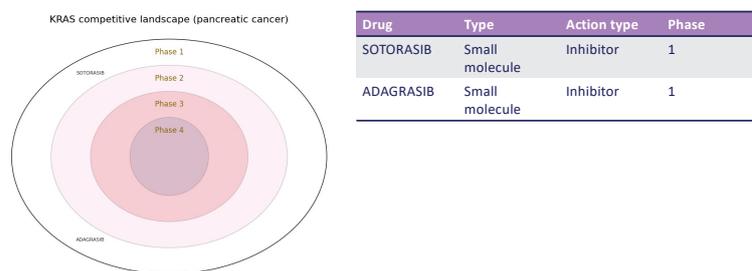


Figure 5.9: Gene essentiality in human organs. From [27].

An overview of the competitive landscape is presented in two slides, as illustrated in Figure 5.10, which provides insights into existing drugs acting on the target of interest. In the first slide, the Agent produces a concentric plot to indicate the current phase of development for existing drugs, using data from Open Targets. A table with details about the drugs, namely their name, type, action type, and phase, is placed next to it. The slide contains data only about drugs that target the gene in a specific disease (pancreatic cancer, in this case). On the other hand, the target dossier PDF provides information on all existing drugs acting on the target, considering various types of cancers. There are only two drugs targeting KRAS for pancreatic cancer, namely Sotorasib and Adagrasib. However, they are still in development and are currently in phase 1, the stage that focuses on gathering short-term safety and pharmacological data by administering the drug to a small group of patients [70]. In the second slide, the molecular structure of the drugs is displayed, and the reader is directed to the PDF for more information about the molecules, such as their description, mechanism of action, and toxicity.

## Known drugs targeting KRAS

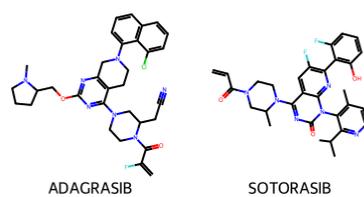


Source: Open Targets

Drugs full list available on the target dossier PDF.



## Known drugs targeting KRAS



\*Molecules information in the PDF.

Source: DrugBank and NCBI PubChem Compound



Figure 5.10: Existing drugs acting on the target. From [27].

The last slides of the presentation include a SWOT analysis that highlights the potential strengths, weaknesses, opportunities, and threats of targeting the gene in the cancer of interest. It is illustrated in Figure 5.11. The analysis is performed by the LLM through prompt engineering. The model receives a prompt with very specific instructions on how to create a SWOT analysis, including a list of questions that it should answer. Additionally, all the relevant information generated earlier (target characteristics, mutations, essentiality, etc.) is added to the prompt. The LLM then follows a specified format to generate the output analysis. This is the most challenging task, as it requires high-level reasoning and in-depth knowledge of drug discovery and development.

## SWOT analysis



➤ Strengths:

- 1. Significant target: KRAS is a frequently mutated oncogene in pancreatic cancer, making it a significant target for drug development.
- 2. Solid foundation: The availability of KRAS mutation data and the understanding of its role in pancreatic cancer progression provide a solid foundation for drug development efforts.
- 3. Central role: KRAS signaling plays a central role in pancreatic cancer progression, making it an attractive target for therapeutic intervention.

➤ Threats:

- 1. Complexity: Developing a drug for KRAS is a complex and challenging process due to the presence of a desmoplastic tumor microenvironment and the lack of effective biomarkers to guide treatment.
- 2. Competition: Several companies are already working on KRAS-targeted therapies, increasing the competition in this area.
- 3. Mortality rate: The high mortality rate of pancreatic cancer and the limited market size due to the low survival rate pose significant threats to drug development efforts.

OPM

## SWOT analysis



➤ Weaknesses:

- 1. Current standard: The modest beneficial outcome of gemcitabine, the current standard of care, necessitates the exploration of combination therapies and alternative treatment options.
- 2. Complexity: The complex nature of pancreatic cancer and the presence of a desmoplastic tumor microenvironment increase the risk of drug development efforts.
- 3. Lack of understanding: The absence of a clear understanding of the mechanisms underlying KRAS-driven tumor progression poses challenges in developing effective KRAS-targeted therapies.

➤ Opportunities:

- 1. Precision medicine: Utilizing precision medicine, such as BRCA mutations, MSI-H/dMMR, and NTRK fusions, can make drug development efforts more efficient by focusing on specific patient populations.
- 2. Therapeutic areas: The potential therapeutic area segments, such as combination therapies and immunotherapy, offer opportunities for expanding the scope of KRAS-targeted drug development.
- 3. Advancements: Advancements in technology, such as CRISPR-Cas9 gene editing and RNA interference, offer opportunities for developing novel KRAS-targeted therapies.

OPM

Figure 5.11: SWOT analysis for the target in pancreatic cancer. From [27].

# Chapter 6

## Discussion

### 6.1 Conclusions

In this thesis, we developed an LLM-based Agent system to support the target identification and validation step in the drug discovery process, focusing on pancreatic cancer due to its limited therapeutic options and low survival rates. The brain component of the Agent is the LLM, which should exhibit high performance and efficiency, ensuring accurately generated outputs and rapid response times. Among the tested models, Mistral-7B offered the optimal trade-off between performance and efficiency, generating high-quality answers in a short amount of time. However, LLMs are trained on massive amounts of general data and may lack in-depth knowledge in specialized domains, such as drug discovery or the biomedical field. This could lead to hallucinations, i.e., the generation of factually incorrect outputs, highlighting the need for adaptation techniques. Our results confirm the need for specialized models for drug discovery. To this end, Mistral-7B was fine-tuned using abstracts from the pancreatic cancer biomedical literature to augment its domain-specific knowledge, with a particular emphasis on pancreatic cancer research. The fine-tuned Mistral-7B was evaluated on a set of questions and showed improved knowledge related to drug discovery with respect to

its base version. However, fine-tuning presents different challenges and limitations, including catastrophic forgetting and the need for high computational resources. In biomedical domains, where new discoveries are frequent and up-to-date information is crucial, fine-tuning becomes impractical. A more efficient adaptation technique is RAG, which requires the maintenance of a vector database to store documents and embeddings that can be easily updated with new knowledge. Two RAG systems were developed and compared. Naive RAG relies solely on semantic search, whereas our advanced RAG system integrates hybrid search (semantic and keyword search) along with a reranking model to ensure that the most relevant documents are ranked highest. Their comparison shows that the retrieval improvements in the advanced RAG process dramatically increase the quality of the responses. Finally, an LLM-based Agent system was implemented, integrating advanced RAG and the use of external tools to access databases and run Python code. The Agent generates an automatic target dossier, which contains all the relevant information to assess the fitness of a gene as a potential target for pancreatic cancer. This system can be crucial, as manually creating a target dossier for each gene requires weeks of work. It would accelerate the drug discovery process, which is of particular importance in highly lethal cancers, such as pancreatic cancer, where new treatment alternatives are urgently needed.

## 6.2 Perspectives

While the Agent system presented in this thesis can significantly help in supporting the drug discovery process, there are several opportunities to improve and extend the system with additional tools and functionalities.

In this thesis, we opted for a relatively small LLM, specifically Mistral-7B. Despite its efficiency, the model presents some limitations, in particular in complex reasoning tasks. A larger model, especially one designed for advanced reasoning, could lead to more accurate results, particularly beneficial

for tasks such as SWOT analysis which require more critical evaluations. A promising approach to overcoming this limitation is Chain-of-Thought fine-tuning. This method enhances the reasoning capabilities of smaller LMs by fine-tuning them on datasets that contain CoT rationales. This improves their ability to tackle complex reasoning tasks, particularly by enhancing their zero-shot generalization performance [54].

The Agent system can generate an automatic target dossier for each gene but is currently specialized only in pancreatic cancer. A natural extension would be to generalize it to other types of tumors, such as lung and breast cancers, which are responsible for many cancer-related deaths. The target dossier is a valuable tool also for other diseases. Therefore, our system can be adapted for other disease areas such immunitary, neurodegenerative, and so on. Especially for rare diseases, having all information at a glance will be highly beneficial, and an extended version of our work can be critical.

Moreover, the number of tools that the Agent utilizes could be expanded, allowing it to connect to many more databases, as well as other models. For instance, the Agent could connect to AlphaFold [50], an AI system developed by DeepMind to predict the 3D structures of proteins. In particular, the introduction of AlphaFold 3 [1] has the potential to be highly beneficial for drug discovery. Indeed, it can predict the structures of protein- molecule complexes, facilitating the identification and design of new molecules that could lead to successful new therapies [20]. Another valuable integration could involve incorporating the model presented in [12], which predicts the pathogenicity of genetic variants, to understand which mutations might be deleterious and prioritize specific targets.

The emergence of multimodal machine learning [59], a research field focused on developing models capable of integrating multiple modalities such as text, image, and audio, has created new opportunities for advancements in drug discovery. These models have shown the advantages of combining structured and unstructured knowledge in different tasks, including predicting

drug–target and drug–drug interactions [61]. Moreover, multimodal machine learning can lead to more accurate molecular property prediction thanks to the combination of three different modalities of information: SMILES-encoded vectors, ECFP fingerprints, and molecular graphs [60]. Furthermore, these models can help in predicting protein-ligand binding affinity by leveraging a Transformer encoder to extract sequence features from both the protein and its binding pocket while integrating graph isomorphism networks to capture ligand-specific features [87]. The integration of these models could expand the Agent’s application in drug discovery beyond target validation, extending its use to lead generation, the process of identifying a compound that could have an effective action against a specific target.

Finally, the Agent could be developed into an application, serving as an automated tool for biologists. This would involve creating both a front-end and back-end to ensure the system is functional for real-world use.

In recent years, AI has shown its potential for application in the medical field. It has facilitated a deeper understanding of disease mechanisms and has sped up patient management through data-driven techniques, which have led to predictive modeling and personalized medicine approaches [35]. New advancements in AI have the potential to revolutionize healthcare, improving patients’ quality of health. In this line, our work is a helpful tool to speed up the research of new therapies and, therefore, extend the therapeutic landscape for a patient.

# List of Acronyms

- **BLAST** - Basic Local Alignment Search Tool
- **DL** - Deep Learning
- **DLEPS** - Deep Learning-based Efficacy Prediction System
- **LLM** - Large Language Model
- **LoRA** - Low Rank Adaptation
- **NCBI** - National Center for Biotechnology Information
- **OOV** - Out-Of-Vocabulary
- **PDAC** - Pancreatic Ductal Adenocarcinoma
- **PEFT** - Parameter-Efficient Fine-Tuning
- **PMC** - PubMed Central
- **RAG** - Retrieval-Augmented Generation
- **SWOT** - Strengths, Weaknesses, Opportunities, and Threats

# List of Annexes

- **Annex A: SwiftDossier: Tailored Automatic Dossier for Drug Discovery with LLMs and Agents.**

This article discusses the development of an LLM Agent system to create an automatic target dossier.

- **Annex B: Integrating Large Language Models for Genetic Variant Classification.**

This article discusses the integration of LLMs to create a system for variant classification.

# Bibliography

- [1] J. Abramson, J. Adler, J. Dunger, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. DOI: 10.1038/s41586-024-07487-w.
- [2] A. A. Aleissae, A. Kumar, R. M. Anwer, S. Khan, H. Cholakkal, G.-S. Xia, and F. S. Khan. Transformers in remote sensing: a survey, 2023. arXiv: 2209.01206 [cs.CV]. URL: <https://arxiv.org/abs/2209.01206>.
- [3] F. Almeida and G. Xexéo. Word embeddings: a survey, 2023. arXiv: 1901.09069 [cs.CL]. URL: <https://arxiv.org/abs/1901.09069>.
- [4] H. Askar, E. Elgeldawi, H. Aboul Ella, Y. A. M. M. Elshaier, M. M. Gomaa, and A. E. Hassanien. Deep learning in drug discovery: an integrative review and future challenges. *Artificial Intelligence Review*, 56:5975–6037, 2023. DOI: 10.1007/s10462-022-10306-1.
- [5] T. H. P. Atlas. The human protein atlas. URL: <https://www.proteinatlas.org>.
- [6] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, B. Hui, L. Ji, M. Li, J. Lin, R. Lin, D. Liu, G. Liu, C. Lu, K. Lu, J. Ma, R. Men, X. Ren, X. Ren, C. Tan, S. Tan, J. Tu, P. Wang, S. Wang, W. Wang, S. Wu, B. Xu, J. Xu, A. Yang, H. Yang, J. Yang, S. Yang, Y. Yao, B. Yu, H. Yuan, Z. Yuan, J. Zhang, X. Zhang,

- Y. Zhang, Z. Zhang, C. Zhou, J. Zhou, X. Zhou, and T. Zhu. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.
- [7] A. Balaguer, V. Benara, R. L. de Freitas Cunha, R. de M. Estevão Filho, T. Hendry, D. Holstein, J. Marsman, N. Mecklenburg, S. Malvar, L. O. Nunes, R. Padilha, M. Sharp, B. Silva, S. Sharma, V. Aski, and R. Chandra. Rag vs fine-tuning: pipelines, tradeoffs, and a case study on agriculture, 2024. arXiv: 2401.08406 [cs.CL]. URL: <https://arxiv.org/abs/2401.08406>.
- [8] R. Barker. A flexible blueprint for the future of drug development. *The Lancet*, 375(9712):357–359, 2010.
- [9] M. Berquez, A. L. Li, M. A. Luy, A. C. Venida, T. O’Loughlin, G. Rademaker, A. Barpanda, J. Hu, J. Yano, A. Wiita, L. A. Gilbert, P. M. Bruno, and R. M. Perera. A multi-subunit autophagic capture complex facilitates degradation of er stalled mhc-i in pancreatic cancer. *bioRxiv [Preprint]*, 2024. DOI: 10.1101/2024.10.27.620516. URL: <https://doi.org/10.1101/2024.10.27.620516>.
- [10] BioAgilytix. Discovery phase in drug development, 2025. URL: <https://www.bioagilytix.com/solutions/phases/discovery-phase-drug-development>.
- [11] A. Blanco-González, A. Cabezón, A. Seco-González, D. Conde-Torres, P. Antelo-Riveiro, Á. Piñeiro, and R. Garcia-Fandino. The role of ai in drug discovery: challenges, opportunities, and strategies. *Pharmaceuticals*, 16(6), 2023. DOI: 10.3390/ph16060891.
- [12] Y. Boulaimen, G. Fossi, L. Outemzabet, N. Jeanray, O. Levenets, S. Gerart, S. Vachenc, S. Raieli, and J. Giemza. Integrating large language models for genetic variant classification, 2024. arXiv: 2411.05055 [q-bio.GN]. URL: <https://arxiv.org/abs/2411.05055>.
- [13] H. Chase. LangChain, October 2022. URL: <https://github.com/langchain-ai/langchain>.

- [14] P.-C. Chen, H. Tsai, S. Bhojanapalli, H. W. Chung, Y.-W. Chang, and C.-S. Ferng. A simple and effective positional encoding for transformers, 2021. arXiv: 2104.08698 [cs.CL]. URL: <https://arxiv.org/abs/2104.08698>.
- [15] C. Christophe, P. K. Kanithi, P. Munjal, T. Raha, N. Hayat, R. Rajan, A. Al-Mahrooqi, A. Gupta, M. U. Salman, G. Gosal, B. Kanakiya, C. Chen, N. Vassilieva, B. B. Amor, M. A. Pimentel, and S. Khan. Med42 – evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches, 2024. arXiv: 2404.14779 [cs.CL]. URL: <https://arxiv.org/abs/2404.14779>.
- [16] Chroma. Chroma database. URL: <https://www.trychroma.com>.
- [17] T. Conroy, F. Desseigne, M. Ychou, O. Bouché, R. Guimbaud, Y. Bécouarn, A. Adenis, J. Raoul, S. Gourgou-Bourgade, C. de la Fouchardière, J. Bennouna, J. Bachet, F. Khemissa-Akouz, D. Péré-Vergé, C. Delbaldo, E. Assenat, B. Chauffert, P. Michel, C. Montoto-Grillot, M. Ducreux, G. T. D. of Unicancer, and P. Intergroup. Folfirinox versus gemcitabine for metastatic pancreatic cancer. *The New England Journal of Medicine*, 364(19):1817–1825, 2011. DOI: 10.1056/NEJMoa1011923.
- [18] T. U. Consortium. Uniprot: the universal protein knowledgebase in 2025. *Nucleic Acids Research*, 53(D1):D609–D617, 2024. ISSN: 1362-4962. DOI: 10.1093/nar/gkae1010. eprint: <https://academic.oup.com/nar/article-pdf/53/D1/D609/60719276/gkae1010.pdf>. URL: <https://doi.org/10.1093/nar/gkae1010>.
- [19] P. Darji, J. Patel, B. Patel, V. Khatri, P. I. J. Fnu, and S. Nalla. Comprehensive review on drug discovery and development process, 2025.
- [20] D. Desai, S. Kantliwala, J. Vybhavi, R. Ravi, H. Patel, and J. Patel. Review of alphafold 3: transformative advances in drug design and therapeutics. *Cureus*, 16(7), 2024. DOI: 10.7759/cureus.63646.

- [21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: pre-training of deep bidirectional transformers for language understanding, 2019. arXiv: 1810.04805 [cs.CL]. URL: <https://arxiv.org/abs/1810.04805>.
- [22] J. Dilly, M. Hoffman, L. Abbassi, Z. Li, F. Paradiso, B. Parent, C. Hennessey, A. Jordan, M. Morgado, S. Dasgupta, G. Uribe, A. Yang, K. Kapner, F. Hambitzer, L. Qiang, H. Feng, J. Geisberg, J. Wang, K. Evans, H. Lyu, A. Schalck, N. Feng, A. Lopez, C. Bristow, M. Kim, K. Rajapakshe, V. Bahrambeigi, J. Roth, K. Garg, P. Guerrero, B. Stanger, S. Cristea, S. Lowe, T. Baslan, E. Van Allen, J. Mancias, E. Chan, A. Anderson, Y. Katlinskaya, A. Shalek, D. Hong, S. Pant, J. Hallin, K. Anderes, P. Olson, T. Heffernan, S. Chugh, J. Christensen, A. Maitra, B. Wolpin, S. Raghavan, J. Nowak, P. Winter, S. Dougan, and A. Aguirre. Mechanisms of resistance to oncogenic kras inhibition in pancreatic cancer. *Cancer discovery*, 14(11):2135–2161, 2024. ISSN: 2159-8274.
- [23] M. Q. Ding, L. Chen, G. F. Cooper, J. D. Young, and X. Lu. Precision oncology beyond targeted therapy: combining omics data with machine learning matches the majority of cancer cells to effective therapeutics. *Molecular Cancer Research*, 16(2):269–278, 2018. DOI: 10.1158/1541-7786.MCR-17-0378.
- [24] Q. Dong, L. Li, D. Dai, C. Zheng, J. Ma, R. Li, H. Xia, J. Xu, Z. Wu, T. Liu, B. Chang, X. Sun, L. Li, and Z. Sui. A survey on in-context learning, 2024. arXiv: 2301.00234 [cs.CL]. URL: <https://arxiv.org/abs/2301.00234>.
- [25] G. A. FitzGerald. Re-engineering drug discovery and development. *LDI Issue Brief*, 17(2):1–4, 2011.
- [26] N. C. for Biotechnology Information. Pubmed central (pmc). URL: <https://pmc.ncbi.nlm.nih.gov>.

- [27] G. Fossi, Y. Boulaimen, L. Outemzabet, N. Jeanray, S. Gerart, S. Vachenc, J. Giemza, and S. Raieli. Swift dossier: tailored automatic dossier for drug discovery with llms and agents, 2024. arXiv: 2409.15817 [cs.AI]. URL: <https://arxiv.org/abs/2409.15817>.
- [28] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, and H. Wang. Retrieval-augmented generation for large language models: a survey, 2024. arXiv: 2312.10997 [cs.CL]. URL: <https://arxiv.org/abs/2312.10997>.
- [29] gathnex. Gath\_baize dataset. URL: [https://huggingface.co/datasets/gathnex/Gath\\_baize](https://huggingface.co/datasets/gathnex/Gath_baize).
- [30] J. Gillson, Y. Ramaswamy, G. Singh, A. A. Gorfe, N. Pavlakis, J. Samra, A. Mittal, and S. Sahni. Small molecule kras inhibitors: the future for targeted pancreatic cancer therapy? *Cancers*, 12(5), 2020. ISSN: 2072-6694. DOI: 10.3390/cancers12051341. URL: <https://www.mdpi.com/2072-6694/12/5/1341>.
- [31] S. Gupta, R. Ranjan, and S. N. Singh. A comprehensive survey of retrieval-augmented generation (rag): evolution, current landscape and future directions, 2024. arXiv: 2410.12837 [cs.CL]. URL: <https://arxiv.org/abs/2410.12837>.
- [32] M. U. Hadi, Q. Al-Tashi, R. Qureshi, A. Shah, A. Muneer, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili. A survey on large language models: applications, challenges, limitations, and practical usage, 2022.
- [33] C. J. Halbrook, C. A. Lyssiotis, M. P. di Magliano, and A. Maitra. Pancreatic cancer: advances and challenges. *Cell*, 186(8):1729–1754, 2023.
- [34] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu, Z. Yang, Y. Zhang, and D. Tao. A survey on

- vision transformer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(1):87–110, 2023. ISSN: 1939-3539. DOI: 10.1109/tpami.2022.3152247. URL: <http://dx.doi.org/10.1109/TPAMI.2022.3152247>.
- [35] M. G. Hanna, L. Pantanowitz, R. Dash, J. H. Harrison, M. Deebajah, J. Pantanowitz, and H. H. Rashidi. Future of artificial intelligence—machine learning trends in pathology and medicine. *Modern Pathology*, 38(4):100705, 2025. ISSN: 0893-3952. DOI: <https://doi.org/10.1016/j.modpat.2025.100705>. URL: <https://www.sciencedirect.com/science/article/pii/S0893395225000018>.
- [36] I. V. Hinkson, B. Madej, and E. A. Stahlberg. Accelerating therapeutics for opportunities in medicine: a paradigm shift in drug discovery. *Frontiers in pharmacology*, 11:770, 2020.
- [37] N. Hiraoka, Y. Ino, S. Hori, R. Yamazaki-Itoh, C. Naito, M. Shimasaki, M. Esaki, S. Nara, Y. Kishi, K. Shimada, N. Nakamura, T. Torigoe, and Y. Heike. Expression of classical human leukocyte antigen class i antigens, hla-e and hla-g, is adversely prognostic in pancreatic cancer patients. *Cancer Science*, 111(8):3057–3070, 2020. DOI: 10.1111/cas.14514.
- [38] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: low-rank adaptation of large language models, 2021. arXiv: 2106.09685 [cs.CL]. URL: <https://arxiv.org/abs/2106.09685>.
- [39] H.-f. Hu, Z. Ye, Y. Qin, X.-j. Y. Xiao-wu Xu, and Q.-f. Z. S.-r. Ji. Mutations in key driver genes of pancreatic cancer: molecularly targeted therapies and other clinical implications, 2021. DOI: 10.1038/s41401-020-00584-2.
- [40] J. Huang, L. Cui, A. Wang, C. Yang, X. Liao, L. Song, J. Yao, and J. Su. Mitigating catastrophic forgetting in large language models with

- self-synthesized rehearsal, 2024. arXiv: 2403.01244 [cs.CL]. URL: <https://arxiv.org/abs/2403.01244>.
- [41] L. Huang, W. Yu, W. Ma, W. Zhong, Z. Feng, H. Wang, Q. Chen, W. Peng, X. Feng, B. Qin, and T. Liu. A survey on hallucination in large language models: principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025.
- [42] Hugging Face. Bge-base-en. URL: <https://huggingface.co/BAAI/bge-base-en>.
- [43] Hugging Face. Mxbai-rerank-large. URL: <https://huggingface.co/mixedbread-ai/mxbai-rerank-large-v1>.
- [44] S. Islam, H. Elmekki, A. Elsebai, J. Bentahar, N. Drawel, G. Rjoub, and W. Pedrycz. A comprehensive survey on applications of transformers for deep learning tasks, 2023. arXiv: 2306.07303 [cs.LG]. URL: <https://arxiv.org/abs/2306.07303>.
- [45] M. R. J, K. VM, H. Warriar, and Y. Gupta. Fine tuning llm for enterprise: practical guidelines and recommendations, 2024. arXiv: 2404.10779 [cs.SE]. URL: <https://arxiv.org/abs/2404.10779>.
- [46] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mistral 7b, 2023. arXiv: 2310.06825 [cs.CL]. URL: <https://arxiv.org/abs/2310.06825>.
- [47] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. de las Casas, E. B. Hanna, F. Bressand, G. Lengyel, G. Bour, G. Lample, L. R. Lavaud, L. Saulnier, M.-A. Lachaux, P. Stock, S. Subramanian, S. Yang, S. Antoniak, T. L. Scao, T. Gervet, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed. Mixtral of experts, 2024. arXiv: 2401.04088 [cs.LG]. URL: <https://arxiv.org/abs/2401.04088>.

- [48] J. Jiao, L. Ruan, C. S. Cheng, F. Wang, P. Yang, and Z. Chen. Paired protein kinases *prkci-ripk2* promote pancreatic cancer growth and metastasis via enhancing *nf-kb/jnk/erk* phosphorylation. *Molecular Medicine*, 29(1):47, 2023. DOI: 10.1186/s10020-023-00648-z.
- [49] M. Juhas, L. Eberl, and J. I. Glass. Essence of life: essential genes of minimal genomes. *Trends in Cell Biology*, 21(10):562–568, 2011. ISSN: 0962-8924. DOI: <https://doi.org/10.1016/j.tcb.2011.07.005>. URL: <https://www.sciencedirect.com/science/article/pii/S0962892411001449>.
- [50] J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. DOI: 10.1038/s41586-021-03819-2.
- [51] F. June. Model quantization 1: basic concepts, 2023. URL: [https://medium.com/@florian\\_algo/model-quantization-1-basic-concepts-860547ec6aa9](https://medium.com/@florian_algo/model-quantization-1-basic-concepts-860547ec6aa9).
- [52] H. JX, Z. CF, C. WB, L. QC, L. QW, L. YY, and G. F. Pancreatic cancer: a review of epidemiology, trend, and risk factors, 2021. DOI: 10.3748/wjg.v27.i27.4298.
- [53] T. Kamisawa, L. D. Wood, T. Itoi, and K. Takaori. Pancreatic cancer, 2016. DOI: 10.1016/S0140-6736(16)00141-0.
- [54] S. Kim, S. J. Joo, D. Kim, J. Jang, S. Ye, J. Shin, and M. Seo. The cot collection: improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning, 2023. arXiv: 2305.14045 [cs.CL]. URL: <https://arxiv.org/abs/2305.14045>.
- [55] J. Kleeff, M. Korc, M. Apte, C. L. Vecchia, C. D. Johnson, A. V. Biankin, R. E. Neale, M. Tempero, D. A. Tuveson, R. H. Hruban, and J. P. Neoptolemos. Pancreatic cancer, 2016. DOI: 10.1038/nrdp.2016.22.

- [56] D. Kumar, A. Kumar, S. Agarwal, and P. Harshangi. Fine-tuning, quantization, and llms: navigating unintended outcomes, 2024. arXiv: 2404.04392 [cs.CR]. URL: <https://arxiv.org/abs/2404.04392>.
- [57] LangChain. How to split text based on semantic similarity. URL: [https://python.langchain.com/docs/how\\_to/semantic-chunker/](https://python.langchain.com/docs/how_to/semantic-chunker/).
- [58] LangChain. Parentdocumentretriever. URL: [https://python.langchain.com/api\\_reference/langchain/retrievers/langchain.retrievers.parent\\_document\\_retriever.ParentDocumentRetriever.html](https://python.langchain.com/api_reference/langchain/retrievers/langchain.retrievers.parent_document_retriever.ParentDocumentRetriever.html).
- [59] P. P. Liang, A. Zadeh, and L.-P. Morency. Foundations and trends in multimodal machine learning: principles, challenges, and open questions, 2023. arXiv: 2209.03430 [cs.LG]. URL: <https://arxiv.org/abs/2209.03430>.
- [60] X. Lu, L. Xie, L. Xu, R. Mao, X. Xu, and S. Chang. Multimodal fused deep learning for drug property prediction: integrating chemical language and molecular graph. *Computational and Structural Biotechnology Journal*, 23:1666–1679, 2024. ISSN: 2001-0370. DOI: <https://doi.org/10.1016/j.csbj.2024.04.030>. URL: <https://www.sciencedirect.com/science/article/pii/S2001037024001132>.
- [61] Y. Luo, X. Y. Liu, K. Yang, K. Huang, M. Hong, J. Zhang, Y. Wu, and Z. Nie. Toward unified ai drug discovery with multimodal knowledge. *Health Data Science*, 4, 2024. DOI: 10.34133/hds.0113.
- [62] J. Lv, Y. Liu, S. Mo, Y. Zhou, F. Chen, F. Cheng, C. Li, D. Saimi, M. Liu, H. Zhang, K. Tang, J. Ma, Z. Wang, Q. Zhu, W. M. Tong, and B. Huang. Gasdermin e mediates resistance of pancreatic adenocarcinoma to enzymatic digestion through a ybx1-mucin pathway. *Nature Cell Biology*, 24(3):364–372, 2022. DOI: 10.1038/s41556-022-00857-4.

- [63] Y. Mao, Y. Ge, Y. Fan, W. Xu, Y. Mi, Z. Hu, and Y. Gao. A survey on lora of large language models. *Frontiers of Computer Science*, 19(7), 2024. ISSN: 2095-2236. DOI: 10 . 1007 / s11704 - 024 - 40663 - 9. URL: <http://dx.doi.org/10.1007/s11704-024-40663-9>.
- [64] S. J. Mielke, Z. Alyafeai, E. Salesky, C. Raffel, M. Dey, M. Gallé, A. Raja, C. Si, W. Y. Lee, B. Sagot, and S. Tan. Between words and characters: a brief history of open-vocabulary modeling and tokenization in nlp, 2021. arXiv: 2112.10508 [cs.CL]. URL: <https://arxiv.org/abs/2112.10508>.
- [65] T. Mikolov, I. Sutskever, A. Deoras, H.-S. Le, S. Kombrink, and J. Černocký. Subword language modeling with neural networks. In 2012.
- [66] S. Minaee, T. Mikolov, N. Nikzad, M. Chenaghlu, R. Socher, X. Amatriain, and J. Gao. Large language models: a survey, 2024. arXiv: 2402.06196 [cs.CL]. URL: <https://arxiv.org/abs/2402.06196>.
- [67] A. Mitra, L. D. Corro, S. Mahajan, A. Cudas, C. Simoes, S. Agrawal, X. Chen, A. Razdaibiedina, E. Jones, K. Aggarwal, H. Palangi, G. Zheng, C. Rosset, H. Khanpour, and A. Awadallah. Orca 2: teaching small language models how to reason, 2023. arXiv: 2311.11045 [cs.AI].
- [68] J. Mizrahi, R. Surana, J. Valle, and R. Shroff. Pancreatic cancer. *Lancet*, 395(10242):2008–2020, 2020. DOI: 10 . 1016 / S0140 - 6736 (20 ) 30974-0.
- [69] R. C. Mohs and N. H. Greig. Drug discovery and development: role of basic biological research. *Alzheimer's Dementia: Translational Research Clinical Interventions*, 3(4):651–657, 2017. ISSN: 2352-8737. DOI: <https://doi.org/10.1016/j.trci.2017.10.005>. URL: <https://www.sciencedirect.com/science/article/pii/S2352873717300653>.

- [70] J. Muglia and J. DiGiovanna. Phase 1 clinical trials. *J Cutan Med Surg*, 2(4):236–241, 1998. DOI: 10.1177/120347549800200413.
- [71] V. B. Parthasarathy, A. Zafar, A. Khan, and A. Shahid. The ultimate guide to fine-tuning llms from basics to breakthroughs: an exhaustive review of technologies, research, best practices, applied research challenges and opportunities, 2024. arXiv: 2408.13296 [cs.LG]. URL: <https://arxiv.org/abs/2408.13296>.
- [72] K. E. Poruk, D. Z. Gay, K. Brown, J. D. Mulvihill, K. M. Boucher, C. L. Scaife, M. A. Firpo, and S. J. Mulvihill. The clinical utility of ca 19-9 in pancreatic adenocarcinoma: diagnostic and prognostic updates. *Current Molecular Medicine*, 13(3):340–351, 2013. DOI: 10.2174/1566524011313030003.
- [73] A. Pourshams et al. The global, regional, and national burden of pancreatic cancer and its attributable risk factors in 195 countries and territories, 1990–2017: a systematic analysis for the global burden of disease study 2017. *The Lancet Gastroenterology & Hepatology*, 4(12):934–947, 2019.
- [74] R. Pradeep, R. Nogueira, and J. Lin. The expando-mono-duo design pattern for text ranking with pretrained sequence-to-sequence models, 2021. arXiv: 2101.05667 [cs.IR]. URL: <https://arxiv.org/abs/2101.05667>.
- [75] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training, 2018.
- [76] N. Rajaraman, J. Jiao, and K. Ramchandran. Toward a theory of tokenization in llms, 2024. arXiv: 2404.08335 [cs.CL]. URL: <https://arxiv.org/abs/2404.08335>.
- [77] A. R. Sajun, I. Zualkernan, and D. Sankalpa. A historical survey of advances in transformer architectures. *Applied Sciences*, 14(10), 2024.

- ISSN: 2076-3417. DOI: 10.3390/app14104316. URL: <https://www.mdpi.com/2076-3417/14/10/4316>.
- [78] W. Sang, Y. Zhou, H. Chen, C. Yu, L. Dai, Z. Liu, L. Chen, Y. Fang, P. Ma, X. Wu, H. Kong, W. Liao, H. Jiang, J. Qian, D. Wang, and Y. H. Liu. Receptor-interacting protein kinase 2 is an immunotherapy target in pancreatic cancer. *Cancer Discovery*, 14(2):326–347, 2024. DOI: 10.1158/2159-8290.CD-23-0584.
- [79] C. W. Schmidt, V. Reddy, H. Zhang, A. Alameddine, O. Uzan, Y. Pinter, and C. Tanner. Tokenization is more than compression, 2024. arXiv: 2402.18376 [cs.CL]. URL: <https://arxiv.org/abs/2402.18376>.
- [80] M. Schneider, T. Hackert, O. Strobel, and M. W. Büchler. Technical advances in surgery for pancreatic cancer, 2021. DOI: 10.1093/bjs/zxab133.
- [81] M. Shafiq and Z. Gu. Deep residual learning for image recognition: a survey. *Applied Sciences*, 12(18), 2022. ISSN: 2076-3417. DOI: 10.3390/app12188972. URL: <https://www.mdpi.com/2076-3417/12/18/8972>.
- [82] N. Singh, P. Vayer, S. Tanwar, J.-L. Poyet, K. Tsaioun, and B. O. Villoutreix. Drug discovery and development: introduction to the general public and patient groups. *Frontiers in Drug Discovery*, 3, 2023.
- [83] S. Sinha and D. Vohora. Chapter 2 - drug discovery and development: an overview. In *Pharmaceutical Medicine and Translational Clinical Research*, pages 19–32. Academic Press, Boston, 2018. DOI: <https://doi.org/10.1016/B978-0-12-802103-3.00002-X>.
- [84] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, P. Tafti, L. Hussenot, P. G. Sessa, A. Chowdhery, A. Roberts, A. Barua, A. Botev, A. Castro-Ros, A. Slone, A. Héliou, A. Tacchetti, A. Bulanova, A. Paterson,

- B. Tsai, B. Shahriari, C. L. Lan, C. A. Choquette-Choo, C. Crepy, D. Cer, D. Ippolito, D. Reid, E. Buchatskaya, E. Ni, E. Noland, G. Yan, G. Tucker, G.-C. Muraru, G. Rozhdestvenskiy, H. Michalewski, I. Tenney, I. Grishchenko, J. Austin, J. Keeling, J. Labanowski, J.-B. Lespiau, J. Stanway, J. Brennan, J. Chen, J. Ferret, J. Chiu, J. Mao-Jones, K. Lee, K. Yu, K. Millican, L. L. Sjoesund, L. Lee, L. Dixon, M. Reid, M. Mikula, M. Wirth, M. Sharman, N. Chinaev, N. Thain, O. Bachem, O. Chang, O. Wahltinez, P. Bailey, P. Michel, P. Yotov, R. Chaabouni, R. Comanescu, R. Jana, R. Anil, R. McIlroy, R. Liu, R. Mullins, S. L. Smith, S. Borgeaud, S. Girgin, S. Douglas, S. Pandya, S. Shakeri, S. De, T. Klimenko, T. Hennigan, V. Feinberg, W. Stokowiec, Y.-h. Chen, Z. Ahmed, Z. Gong, T. Warkentin, L. Peran, M. Giang, C. Farabet, O. Vinyals, J. Dean, K. Kavukcuoglu, D. Hassabis, Z. Ghahramani, D. Eck, J. Barral, F. Pereira, E. Collins, A. Joulin, N. Fiedel, E. Senter, A. Andreev, and K. Kenealy. Gemma: open models based on gemini research and technology, 2024. arXiv: 2403.08295 [cs.CL]. URL: <https://arxiv.org/abs/2403.08295>.
- [85] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need, 2023. arXiv: 1706.03762 [cs.CL]. URL: <https://arxiv.org/abs/1706.03762>.
- [86] A. S. Verkman. Drug discovery in academia. *American Journal of Physiology-Cell Physiology*, 286(3):C465–C474, 2004. DOI: 10.1152/ajpcell.00397.2003. eprint: <https://doi.org/10.1152/ajpcell.00397.2003>. PMID: 14761879.
- [87] G. Wang, H. Zhang, M. Shao, et al. Deeptgin: a novel hybrid multi-modal approach using transformers and graph isomorphism networks for protein-ligand binding affinity prediction. *Journal of Cheminformatics*, 16:147, 2024. DOI: 10.1186/s13321-024-00938-6.

- [88] H. Wang, S. Hao, H. Dong, S. Zhang, Y. Bao, Z. Yang, and Y. Wu. Offline reinforcement learning for llm multi-step reasoning, 2024. arXiv: 2412.16145 [cs.LG]. URL: <https://arxiv.org/abs/2412.16145>.
- [89] L. Wang, C. Ma, X. Feng, Z. Zhang, H. Yang, J. Zhang, Z. Chen, J. Tang, X. Chen, Y. Lin, W. X. Zhao, Z. Wei, and J. Wen. A survey on large language model based autonomous agents. *Frontiers of Computer Science*, 18(6), 2024. ISSN: 2095-2236. DOI: 10.1007/s11704-024-40231-1. URL: <http://dx.doi.org/10.1007/s11704-024-40231-1>.
- [90] Z. Wang, G. Zhang, K. Yang, N. Shi, W. Zhou, S. Hao, G. Xiong, Y. Li, M. Y. Sim, X. Chen, Q. Zhu, Z. Yang, A. Nik, Q. Liu, C. Lin, S. Wang, R. Liu, W. Chen, K. Xu, D. Liu, Y. Guo, and J. Fu. Interactive natural language processing, 2023. arXiv: 2305.13246 [cs.CL]. URL: <https://arxiv.org/abs/2305.13246>.
- [91] L. Weng. Llm powered autonomous agents, 2023. URL: <https://lilianweng.github.io/posts/2023-06-23-agent/>.
- [92] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush. Huggingface’s transformers: state-of-the-art natural language processing, 2020. arXiv: 1910.03771. URL: <https://arxiv.org/abs/1910.03771>.
- [93] Z. Xi, W. Chen, X. Guo, W. He, Y. Ding, B. Hong, M. Zhang, J. Wang, S. Jin, E. Zhou, R. Zheng, X. Fan, X. Wang, L. Xiong, Y. Zhou, W. Wang, C. Jiang, Y. Zou, X. Liu, Z. Yin, S. Dou, R. Weng, W. Cheng, Q. Zhang, W. Qin, Y. Zheng, X. Qiu, X. Huang, and T. Gui. The rise and potential of large language model based agents: a survey, 2023.

- arXiv: 2309.07864 [cs.AI]. URL: <https://arxiv.org/abs/2309.07864>.
- [94] X. Xie, T. Yu, X. Li, N. Zhang, L. J. Foster, C. Peng, W. Huang, and G. He. Recent advances in targeting the "undruggable" proteins: from drug discovery to clinical trials. *Signal Transduction and Targeted Therapy*, 8(1):335, 2023. DOI: 10.1038/s41392-023-01589-z.
- [95] J. Xu, X. Sun, Z. Zhang, G. Zhao, and J. Lin. Understanding and improving layer normalization, 2019. arXiv: 1911.07013 [cs.LG]. URL: <https://arxiv.org/abs/1911.07013>.
- [96] L. Xu, H. Xie, S.-Z. J. Qin, X. Tao, and F. L. Wang. Parameter-efficient fine-tuning methods for pretrained language models: a critical review and assessment, 2023. arXiv: 2312.12148 [cs.CL]. URL: <https://arxiv.org/abs/2312.12148>.
- [97] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao. React: synergizing reasoning and acting in language models, 2023. arXiv: 2210.03629 [cs.CL]. URL: <https://arxiv.org/abs/2210.03629>.
- [98] A. J. Yepes, Y. You, J. Milczek, S. Laverde, and R. Li. Financial report chunking for effective retrieval augmented generation, 2024. arXiv: 2402.05131 [cs.CL]. URL: <https://arxiv.org/abs/2402.05131>.
- [99] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. A survey of large language models, 2024. arXiv: 2303.18223 [cs.CL]. URL: <https://arxiv.org/abs/2303.18223>.
- [100] Y. Zheng, H. Y. Koh, M. Yang, L. Li, L. T. May, G. I. Webb, S. Pan, and G. Church. Large language models in drug discovery and development: from disease mechanisms to clinical trials, 2024. arXiv:

- 2409.04481 [q-bio.QM]. URL: <https://arxiv.org/abs/2409.04481>.
- [101] Z. Zhong, H. Liu, X. Cui, X. Zhang, and Z. Qin. Mix-of-granularity: optimize the chunking granularity for retrieval-augmented generation, 2025. arXiv: 2406.00456 [cs.LG]. URL: <https://arxiv.org/abs/2406.00456>.
- [102] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models, 2023. arXiv: 2311.07911 [cs.CL]. URL: <https://arxiv.org/abs/2311.07911>.
- [103] J. Zhu, J. Wang, X. Wang, M. Gao, B. Guo, M. Gao, J. Liu, Y. Yu, L. Wang, W. Kong, Y. An, Z. Liu, X. Sun, Z. Huang, H. Zhou, N. Zhang, R. Zheng, and Z. Xie. Prediction of drug efficacy from transcriptional profiles with deep learning. *Nature Biotechnology*, 39(11):1444–1452, 2021. DOI: 10.1038/s41587-021-00946-z.

# Acknowledgements

I would like to express my gratitude to all those who have supported me throughout my academic journey.

I am deeply thankful to my internship supervisor, Salvatore Raieli, for creating a positive and enjoyable atmosphere throughout my internship. His guidance, expertise and passion for LLMs have significantly contributed to my personal and professional growth. I would also like to thank my academic supervisor, Professor Paolo Torroni, for all his advice, support, and assistance during this process and for helping me with various procedures. His dedication is invaluable in creating opportunities for students. Moreover, I am grateful to the rest of the team for their valuable suggestions and ongoing help.

Finally, I would like to express my gratitude to my family and friends for always believing in my abilities, supporting me in every possible way, and providing me with companionship during this journey. I am truly happy to have them as part of my life.

# **Annex A**

**SwiftDossier: Tailored Automatic Dossier for Drug  
Discovery with LLMs and Agents**

# SwiftDossier: Tailored Automatic Dossier for Drug Discovery with LLMs and Agents

Gabriele Fossi<sup>a</sup>, Youssef Boulaimen<sup>a</sup>, Leila Outemzabet<sup>a</sup>, Nathalie Jeanray<sup>a</sup>, Stéphane Gerart<sup>a</sup>, Sébastien Vachenc<sup>a</sup>, Joanna Giemza<sup>a</sup>, Salvatore Raieli<sup>a,b</sup>

<sup>a</sup>*Oncodesign Precision Medicine, 18 rue Jean Mazon, Dijon, 21000, France*

<sup>b</sup>*Corresponding author: sraieli@oncodesign.com*

---

## Abstract

The advancement of artificial intelligence algorithms has expanded their application to several fields such as the biomedical domain. Artificial intelligence systems, including Large Language Models (LLMs), can be particularly advantageous in drug discovery, which is a very long and expensive process. However, LLMs by themselves lack in-depth knowledge about specific domains and can generate factually incorrect information. Moreover, they are not able to perform more complex actions that imply the usage of external tools. Our work is focused on these two issues. Firstly, we show how the implementation of an advanced RAG system can help the LLM to generate more accurate answers to drug-discovery-related questions. The results show that the answers generated by the LLM with the RAG system surpass in quality the answers produced by the model without RAG. Secondly, we show how to create an automatic target dossier using LLMs and incorporating them with external tools that they can use to execute more intricate tasks to gather data such as accessing databases and executing code. The result is a production-ready target dossier containing the acquired information summarized into a PDF and a PowerPoint presentation.

*Keywords:* LLM Agents, Automatic Target Dossier, LLMs in Biology, Drug Discovery

---

## 1. Introduction

Drug discovery is an expensive, lengthy, and high-risk process. The process can cost up to \$ 1-2 billion and takes an average of 10-15 years [19]. The failure rate of a drug candidate in clinical trials reaches 90 % (with a higher rate if preclinical stages are also considered). Among the main causes that can be identified are: lack of clinical efficacy, toxicity, and poor pharmacokinetic and pharmacodynamic properties [9, 8, 5]. In recent years, the development of new artificial intelligence algorithms has been extended to various fields, including biomedical science. The large availability of available data means that the pharmaceutical sector can be a perfect playground for AI [10, 17]. The development of these algorithms promises to revolutionize drug discovery and ameliorate the process from a target to a drug on the market. Drug discovery is a complex, multi-step process; AI looks promising in most steps [4, 1, 18]. At the same time, a cautionary approach is required because the capabilities of an AI model depend heavily on the data with which it is fed. For these promises to be fulfilled, AI approaches must be rigorous and based on quality data [30]. Large Language Models (LLMs) have become the main direction of artificial intelligence in recent years and have shown remarkable generalist skills. LLMs have also shown promising capabilities for different scientific tasks and also in the biomedical field [54, 25]. There has been considerable effort to adapt these models to the needs of the medical and at the same time pharmaceutical fields [11, 50]. One of the main problems with the use of LLMs in the pharmaceutical field (and more broadly in the biomedical field) is the generation of hallucinations by models [32]. LLMs can generate outputs that are inaccurate or

that contain factually incorrect information. At the same time, they can generate extrinsic hallucinations (claims that cannot be verified) such as quotes that cannot be verified. Another problem that affects their use is that these models are generalist by nature and are not often adapted to the domain of interest [34]. In addition, the architecture of LLMs (transformer derivative) does not easily and not expensively allow ingesting new knowledge (as well as making it difficult to correct previously acquired erroneous knowledge ) [29, 27]. The first step in the target discovery process is generally target identification and selection. Once a target has been identified, the target dossier is one of the most widely used tools for evaluating the goodness-of-fit of a target [6, 20]. The target dossier comprehensively presents all the information that experts need in the decision-making process. A target dossier is an essential tool that must present information about the target, its impact on the disease, pros and cons, therapeutic opportunities, and potential competitors. The dossier preparation process requires extensive time in researching and analyzing different sources. Automating it will save considerable time and help in standardization. For appropriate decision-making, the target dossier must contain correct, up-to-date, and verifiable information. For this reason, it is not possible today to use a classical LLM alone for this task.

In this technical report, we show how it is possible to use an LLM to create an automatic target dossier. Our system that employs both Agents and retrieval augmented generation (RAG) allows us to overcome the limitations highlighted above. The system, given a target (gene name) and a pathology of interest, automatically generates a target dossier (a summary PDF document) and a PowerPoint presentation. The system retrieves

up-to-date information from various databases and allows verifiability of the information by annotating the sources used. Our main contributions are:

- we show how a modified pipeline of RAG dramatically increases the quality of response in the biomedical domain.
- we show how using an LLM is possible to create an automatic target dossier. The system accesses article databases, external databases, and tools, executing code and plot autonomously. In addition, it produces a high-quality summary PDF and a ready-to-use PowerPoint presentation.

A complete example of the generated target dossier (both PDF and PowerPoint) is present at <https://github.com/Oncodesign-Precision-Medicine/Automatic-Target-Dossier>

## 2. Related Work

### 2.1. LLM hallucinations

Hallucinations are considered among the most important challenges for LLM deployment in biomedical or drug discovery [54]. Hallucinations can occur when models produce texts that include details, facts, or claims that are either fictional or misleading [32]. In addition, LLMs are capable today of producing content that sounds plausible and based on scientific patterns they learned during training. In some cases, it has been seen that an LLM is capable of generating pseudo-citations that can further confuse the reader. Hallucinations are a real problem, especially for those disciplines where a pattern must necessarily be factually correct (e.g. healthcare) [54]. It is not easy to be able to correct this behavior because there are different types of hallucinations each with probably a different cause [41]. For this reason, several approaches have been developed to mitigate hallucinations in the outputs produced by LLMs [51]. Some techniques are based on the idea of providing the model with the necessary context to be able to answer a question. Retrieval-augmented generation (RAG) allows the LLM to access authoritative, external information, especially when the answer is not present in the parametric memory of the model [46, 43]. In contrast to other methods such as fine-tuning the computational cost is much lower and it is possible to integrate new information easily [40]. However, naive RAG often has limitations (pre-retrieval, retrieval, and post-retrieval issues) [43] so numerous add-ons have evolved [43]. At the same time, finding an effective strategy to create an advanced pipeline of RAGs is an expensive and time-consuming process. One of the sore points of RAG is especially to make sure that all relevant documents are found. In the biomedical field, the exact answer can be directly found in specialized databases. For this in recent times, Agents have been used where complex reasoning is required, thus allowing LLMs to control different tools and query different databases via API [53]. In this work, we use both RAG and Agents to interact with databases.

### 2.2. Domain adaptation and continuous learning with LLMs

LLMs acquire most of their knowledge during pretraining. Once pretraining is over, it is difficult to ingest new knowledge within the model. This can be done by fine-tuning. However, this is an expensive and inefficient process that would have to be repeated every time the model update is to be conducted [37]. In addition, LLM struggles to integrate this new knowledge. Recent studies show that fine-tuning increases the model's tendency to hallucinate [44]. Second, the LLM tends to forget the old knowledge, and there arises at the same time the discrepancy between the remaining old knowledge and the new knowledge with the risk of bias and unintended association [27]. RAG and Agents present a possible solution to these problems. It is not necessary to conduct parameter editing of the LLM but the new knowledge needed is found in real-time and provided in context to the model. Moreover, in several databases the information is rigorously cataloged and annotated, allowing filtering of irrelevant information.

### 2.3. LLM for target dossier

To the best of our knowledge, there is currently no system that creates an automatic target dossier. LLMs have been proposed for various uses in drug discovery pipelines. For example for molecule optimization, drug-drug interaction, molecular property predictions [38, 28, 52]. On the other hand, in several biomedical tasks, generalist LLMs perform poorly [47]. There are, however, studies that show that it is effective to extend the capabilities of models with the use of tools. ToolLLM and HuggingGPT are examples where an LLM is connected with different APIs or can call other models to perform different tasks [33, 31]. GeneGPT connects an LLM with external APIs from the National Center for Biotechnology Information (NCBI) [45]. The system shows superior capabilities in GeneTuring benchmark vs. other generalist or trained models on biomedical tasks. MedAgents is another framework that shows the superior capabilities of an agent-based system for medical tasks [35]. The capabilities of these systems show the goodness of an approach no longer based only on one LLM but with the integration of other tools.

## 3. Materials and methods

### 3.1. Data

A target dossier includes a large variety of information, spanning from insights about a specific target gene to details about the disease, and includes sections in which reasoning is crucial. To this end, several data coming from different sources are needed, in order to have a comprehensive view of the target and allow the LLM to perform high-level reasoning. The data sources and tools employed in this work are listed in Table 1. The specific sources used are specified in both the target dossier PDF and PowerPoint as well. We used reference citations to allow us to track back which sources are used for each section and slide. For example, when the LLM is using the RAG, the identifiers (Pubmed ID) of the article in the context are noted in the document. In the PowerPoint in the notes are stored the

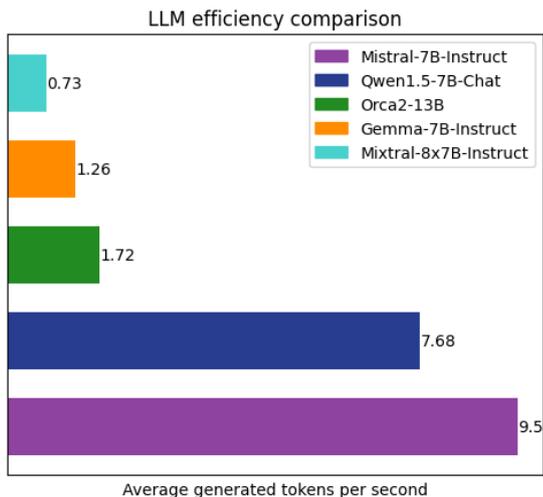


Figure 1: Comparison of LLM efficiency. The height of the bars indicates the average number of generated tokens per second. The value is indicated on the right of the bar.

details of each data used. As an example, when the structure of a protein is used, we note the accession link allowing us to retrieve which information the LLM used.

### 3.2. Models

We decided to use in this work open-source LLMs. Many open-source models differ in architecture and number of parameters [39, 48]. Thus, different LLMs were compared (Mistral-7B-instruct, Qwen-1.5-7B-Chat, Orca 13B, Gemma-7B-instruct, Mistral-8x7B-instruct), focusing on the quality of their answers to a specific set of questions and their efficiency in token generation. Figure 1 shows the results of this analysis. All the assessed models produced good answers, with Mixtral-8x7B-Instruct excelling in this task. However, being a big model, it was slower compared to the other ones. An optimal trade-off between effectiveness and efficiency was Mistral-7B-Instruct [21] which was picked and used in our work.

LLMs contain a substantial amount of information but highly rely on the data that has been used to train them [49]. To expand their internal knowledge by integrating retrieval-based memories, it is possible to leverage Retrieval-Augmented Generation (RAG) [13]. RAG is particularly useful in domain-specific applications, especially in the biomedical domain as described above.

Our RAG framework incorporates reranking and Hybrid Search methodologies, with the embedding model fine-tuned to enhance search effectiveness. Specifically, we utilized *bge-base-en* [36] for this purpose. The model was fine-tuned using a synthetic dataset about pancreatic cancer generated by Qwen 7B [22]. Regarding the reranking model, *mx-bai-rerank-large* was selected. Both these models are relatively small, making them suitable to execute RAG dynamically.

The information that is used for RAG and is provided as context for Mistral 7B is gathered from two different sources: PubMed abstracts and PMC articles. The abstracts are short,

therefore they do not need any preprocessing and can be directly embedded. On the other hand, articles are too long. Thus they are split into chunks using LangChain *SemanticChunker* before the embedding process. The abstracts and articles are embedded using our fine-tuned *bge-base-en* model and stored in a temporary Chroma DB collection. A similarity search between the query and the embeddings is performed, resulting in a list of documents that are reranked and presented to the LLM as the context to generate a text or answer a question. After the termination of the process, the documents and embeddings in the temporary collection are deleted. Besides the content of the abstracts and articles, their PMIDs are extracted too, to reference the specific documents that have been used by the LLM to generate a response. We evaluated our RAG strategy using GPT4 as evaluator [24]. In this work, we compared the answers to 22 drug discovery questions that were written by experts in the domain. We compared three different possibilities: LLM alone (Mistral 7B), naive RAG, and our RAG pipeline.

### 3.3. System Description

The process begins with the user inputting a query that is related to the generation of a target dossier. Both the target and pathology of interest need to be specified. In this work, we focused on pancreatic cancer, which is a deadly disease that needs therapeutic options [26]. The query is directed to the Agent which has different tools at its disposal. These tools allow the Agent to connect to multiple databases and extract the relevant information. They enable the Agent to:

- Execute Python code to connect to the APIs, process data, and generate plots,
- Retrieve images from the databases,
- Generate texts and summaries using the LLM integrated with the RAG system described previously.

Once all the information has been retrieved, it is aggregated and compiled into the target dossier, producing both a PDF and a PowerPoint presentation. The PDF is generated before the PowerPoint since the latter consists of an encapsulated version of the PDF. The system is summarized in Figure 2.

### 3.4. Technical Requirements

This work revolves around three main platforms/applications: HuggingFace [12], LangChain [16] and Chroma DB [55]. HuggingFace is the platform that we used for model loading and usage since it provides a wide variety of pre-trained models that can be easily accessed. LangChain is the framework that we employed to develop the Agent starting from the LLM and for the creation of the tools. An advantage of using LangChain is that it allows an easy integration with HuggingFace. Finally, as a database to implement our RAG system, we used Chroma DB. It allows the storage of the document embeddings and their retrieval during the RAG process. These tools were used through version 3.11.5 of Python. The system was run on 8 Tesla V100 GPUs, each with 32GB of memory.

Source	Description	Web Address
UniProt	A database offers high-quality and freely accessible resource for protein sequence and functional information.	uniprot.org
Human Protein Atlas	A database focused on genome-wide analysis of human proteins	proteinatlas.org
DrugBank	An online database containing information on drugs and drug targets.	go.drugbank.com
Open Targets	A database that uses human genetics and genomics data for drug target identification and prioritization.	opentargets.org
RCSB PDB	An online database containing information on drugs and drug targets.	rcsb.org
cBioPortal	A resource for the exploration of cancer genomics datasets containing information such as mutation frequencies.	cbioportal.org
TCGA Survival	A website that provides analysis of mutations, copy number alterations, etc. Associated with cancer outcome in TCGA.	tcga-survival.com
OGEE	An online database containing information about gene essentiality.	v3.ogee.info
STRING	A database about protein-protein interactions including both direct and indirect associations.	string-db.org
SIGNOR	A repository of annotated causal relationships among human proteins, chemicals of biological relevance, stimuli and phenotypes	signor.uniroma2.it
ESMO	The European Society for Medical Oncology provides oncology information, including guidelines on cancer.	esmo.org
PubChem	A database that contains information about molecules such as chemical structures, chemical and physical properties and biological activities.	pubchem.ncbi.nlm.nih.gov
Gene	An NCBI database that integrates gene information from different species.	ncbi.nlm.nih.gov/gene
PubMed	An NCBI database comprises millions of citations and abstracts about biomedical literature.	pubmed.ncbi.nlm.nih.gov
PMC	An NCBI archive of biomedical and life sciences journal literature.	ncbi.nlm.nih.gov/pmc
BLAST	A tool that finds the similarity between biological sequences	blast.ncbi.nlm.nih.gov/Blast.cgi
DeepTMHMM	A Deep Learning Model for classification and prediction of Transmembrane Topology	dtu.biolib.com/DeepTMHMM
GSEAPy	A Python package to perform Gene Set Enrichment Analysis	github.com/zqfang/GSEAPy

Table 1: List of the sources containing a short description and the web address to access them.

## 4. Results

### 4.1. RAG adaptation to drug discovery domain

To evaluate the effectiveness of RAG, the answers of three different models were compared: the LLM without RAG, the LLM with the base RAG, and the LLM with our RAG system illustrated in Section 3.2. The answers were evaluated by GPT-4 which was specifically prompted to assign a score to each answer according to 4 metrics (faithfulness, relevance, quality, completeness, and correctness). The scores are summarized in the Box Plots of Figure 3. The model with the complete RAG system outperforms the other two models. 75% of its scores are equal to or greater than 4 for every metric. Moreover, its median is 5 and is greater than the median of the other two models which is equal to 4.

One question and the models answers are illustrated in Table 2.

The answer of the model without RAG is entirely incorrect since it fails to understand that pancreatic ductal adenocarcinoma (PDAC) is the subject of the question. Instead, it dis-

cusses the Pacific Data and Automation Corporation which has the same acronym PDAC.

### 4.2. Automatic Target Dossier

In this study, we focused on pancreatic cancer, one of the cancers with the most dismal prognosis. To date, pancreatic cancer presents few therapeutic opportunities and is in dire need of new drugs for treatment [14]. To show the system’s capabilities, we chose KRAS, one of the genes with the greatest impact on the disease and of interest as a potential target [3, 2, 15, 7]. However, the system is generalizable and can therefore be used with any gene and any pathology. Choosing another gene and another pathology, the system will search for the necessary information. The output of the system is the target dossier, generated as two different files: a PDF and a PowerPoint presentation. The table of contents of the PDF is illustrated in Figure 4. It is divided into four main sections. The first one (containing most of the information) comprises details about the target of interest such as its characteristics, expression, and mutations. The second section is related to the cancer for which we are evaluating

<b>Question: "Can you describe the PDAC escape mechanisms?"</b>	
<b>Without RAG</b>	<p>The PDAC (Pacific Data and Automation Corporation) escape mechanisms are a set of security features designed to prevent unauthorized access to the data stored on a PDAC device. These mechanisms include:</p> <ol style="list-style-type: none"> <li>1. Encryption: All data stored on a PDAC device is encrypted using a proprietary algorithm. This ensures that even if an attacker gains access to the device, they will not be able to read the data without the encryption key.</li> <li>2. Authentication: Before accessing any data on a PDAC device, the user must first authenticate themselves using a valid username and password. This ensures that only authorized users can access the data.</li> <li>3. Access Control: Access to data on a PDAC device is controlled using a set of permissions. Each user is assigned a set of permissions that determine what data they can access and what actions they can perform on that data.</li> <li>4. Audit Trail: All actions performed on a PDAC device are logged in an audit trail. This allows administrators to track who accessed the data and when, and to detect any unauthorized access attempts.</li> <li>5. Physical Security: PDAC devices are designed to be physically secure, with features such as tamper-evident seals and secure locking mechanisms to prevent unauthorized access to the device.</li> </ol> <p>Overall, the PDAC escape mechanisms are designed to provide a high level of security for the data stored on a PDAC device, and to prevent unauthorized access to that data.</p>
<b>Base RAG</b>	<p>The provided context discusses two possible escape mechanisms for Pancreatic ductal adenocarcinoma (PDAC):</p> <ol style="list-style-type: none"> <li>1. Gasdermin E (GSDME) and mucin expression: PDAC cells use GSDME to form a barrier that prevents chymotrypsin-mediated destruction. GSDME facilitates the expression of mucin 1 and mucin 13, which form this barrier. Inhibition or knockout of mucin 1 or mucin 13 abrogates orthotopic PDAC growth in NOD-SCID mice.</li> <li>2. Autophagy and Major Histocompatibility Complex class I (MHC-I) sequestration: Autophagy was recently described as a tumor-intrinsic immune escape process during tumor development by sequestering MHC-I inside PDAC cells. This discovery implies that autophagy may be a potential target for limiting immune escape in patients and improving immunotherapy efficiency.</li> </ol> <p>It is important to note that the context does not provide enough information to fully understand the PDAC escape mechanisms or their implications for treatment. Further research and studies are needed to fully understand and address these mechanisms.</p>
<b>New RAG</b>	<p>Based on the provided documents, it appears that pancreatic ductal adenocarcinoma (PDAC) has several escape mechanisms that allow it to evade the immune system and survive in the pancreatic microenvironment. One such mechanism involves the use of gasdermin E (GSDME) to form a barrier that prevents chymotrypsin-mediated destruction. GSDME facilitates the expression of mucin 1 and mucin 13, which form this barrier. Inhibition or knockout of mucin 1 or mucin 13 abrogates orthotopic PDAC growth in NOD-SCID mice. Another mechanism involves the sequestration of Major Histocompatibility Complex class I (MHC-I) inside PDAC cells through autophagy. This process supports tumor progression and may be a potential therapeutic target. Additionally, tumor-derived HLA-G may indicate the mechanism of immune escape and impaired PDAC clinical outcome, especially in early-stage patients. Overall, understanding these escape mechanisms is crucial for the development of more effective immunotherapies for PDAC.</p>

Table 2: Table showing the answers of the three models to the question "Can you describe the PDAC escape mechanisms?"

the impact of the target and is mainly generated by the LLM using the RAG mechanism. The following section focuses on the existing drugs targeting the gene, with particular attention to the drugs used to treat the cancer of interest. This section is significant from the perspective of competition. Finally, the last section encompasses the conclusions and is the most challenging one for the LLM. Indeed, the SWOT analysis requires high-level reasoning and understanding as well as a highly specific prompt. After the generation of the PDF, the information is summarized and compiled into a PowerPoint presentation.

Both PDF and presentation contain references, specifying

which database/tool was used to obtain the information. When available, specific web addresses or IDs are provided, in order to facilitate the reader to access the data and the documents that have been used by the LLM for generation. In the PDF, these specific links can be found at the end in the 'Sources' section. Conversely, in the PowerPoint presentation, they are added in the speaker notes.

Three example slides are illustrated below. Figure 5 shows a table with some characteristics of the target: the similarity between the human protein sequence and the sequences of some animals as well as the protein function. The Agent retrieves

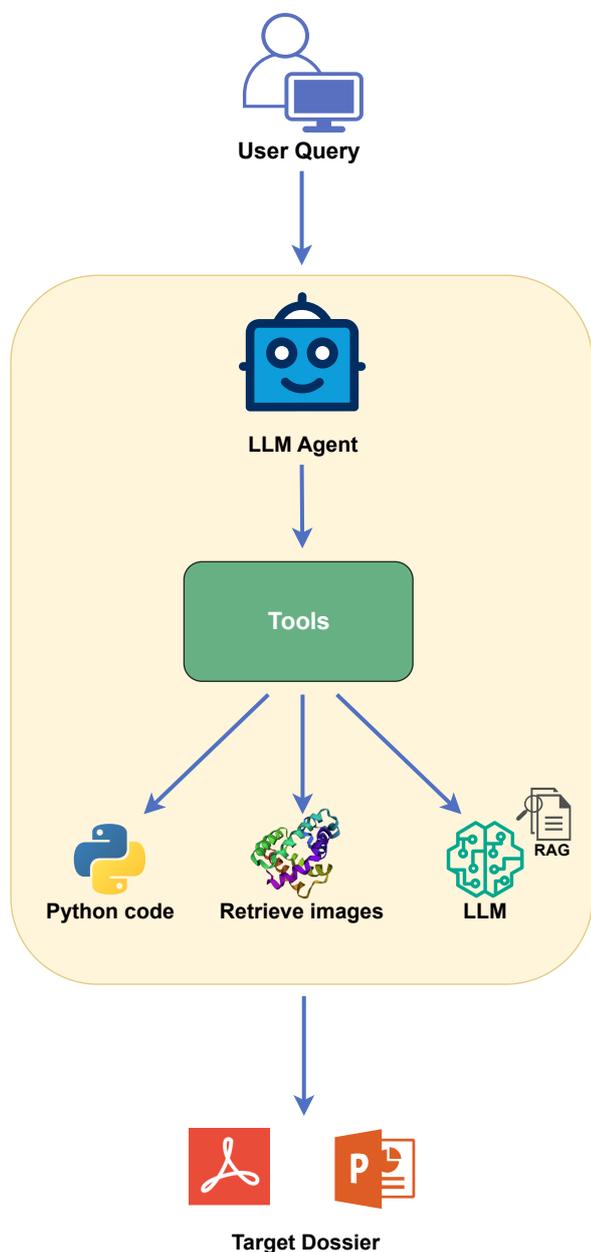


Figure 2: Diagram of the system. The user query is sent to the Agent that uses the available tools to extract information from the databases and generate texts. The information is collected into the target dossier PDF and PowerPoint which are generated and saved.

the sequences from the UniProt API and they are written in FASTA files, a format which is suitable to represent protein sequences. These files are sent to BLAST which returns the alignment match which can be used to compute the similarity. The protein function is extracted from UniProt and is summarized by the LLM in case it exceeds a certain length. This slide shows the system’s ability to retrieve information using external tools, summarize texts using the LLM, and display the data in a table to facilitate readability.

Figure 6 provides insights about the subcellular location of the target in the Human Protein Atlas database. The three pictures illustrating the location are extracted one by one and the names of the represented cell lines are added in the lower-right corner of each image. The antibody used for the analysis is reported in the slide and a legend explaining the colours in the image is added. The image allows the reader to have a clear view of the subcellular location of the target (production-ready). This slide shows the ability of the system to recover images from databases and websites.

Figure 7 illustrates the Agent’s ability to execute Python code. Firstly, the list of the one hundred genes with the highest interaction scores with the target is gathered from the STRING database. Afterward, pathway enrichment analysis is performed using the Python library GSEAPy (based on the previous obtained list from STRING). Both gene lists and gene sets need to be specified. In our case, as gene sets, we considered Gene Ontology Biological Process 2023, Molecular Function 2023, and Cellular Component 2023 as well as KEGG 2021 Human. Lastly, the results of the pathway enrichment analysis are plotted and the graph is added to the target dossier.

## 5. Conclusion

Our work demonstrates two key findings. Firstly, we showed how the employment of RAG can help increase the quality of the answers in drug discovery applications. Indeed, the LLM without RAG lacks knowledge about the biomedical domain and can generate off-topic answers that are unrelated to the subject of the question. Moreover, the exploitation of a more advanced RAG system determines an additional enhancement in the quality of the responses, due to an improved retrieval process. Regarding the generation of the automatic target dossier, we showed how it is possible to integrate the LLM with some tools that allow it to perform more complex actions such as accessing databases, executing Python code, and compiling all the retrieved information into a PDF and a PowerPoint presentation. We show a system capable of creating production-ready documents and presentations and can have a valuable impact on the work in drug discovery.

Toxicity and lack of clinical efficacy are among the main causes of molecule failure. A detailed target dossier eviscerates the role of a gene in pathology, and potential adversarial effects helping in predicting these critical causes of failure. Indeed, in the target dossier, there are sections devoted to both potential toxicity and the role the gene plays in disease physiology and pathogenicity, aiding in decision-making. So, our second finding is that an AI model can generate a consistent document and

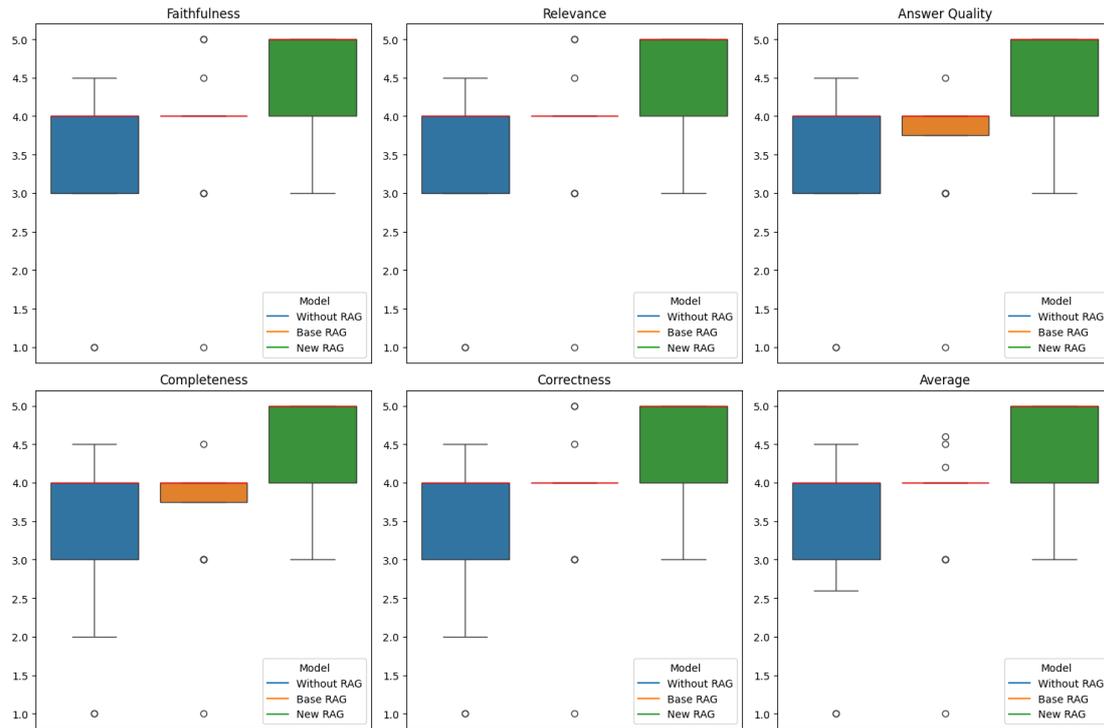


Figure 3: Box Plots summarizing LLM answers evaluation performed by GPT-4. The median is represented by the horizontal red line. The plots show the superior performance of the model with complete RAG.

**Table of contents:**

- 1. Target information..... 3**
  - 1.1 Summary and characteristics..... 3
  - 1.2 Transmembrane Helix Prediction..... 5
  - 1.3 Subcellular location..... 6
  - 1.4 Expression..... 7
  - 1.5 Mutations..... 11
  - 1.6 Glycosylations..... 15
  - 1.7 Gene essentiality..... 16
  - 1.8 Protein-protein interactions..... 19
  - 1.9 Pathway enrichment..... 20
  - 1.10 Signaling Network..... 21
  - 1.11 Role in physiology..... 25
  - 1.12 Role in tumor progression..... 25
  - 1.13 Kaplan-Meier curves..... 26
- 2. Disease information..... 34**
  - 2.1 Disease description..... 34
  - 2.2 Disease statistics..... 35
  - 2.3 ESMO guidelines..... 37
- 3. Competitive landscape..... 43**
  - 3.1 Pancreatic cancer standard of care..... 43
  - 3.2 Pancreatic cancer current therapies..... 44
  - 3.3 Known drugs targeting KRAS..... 45
- 4. Conclusion..... 53**
  - 4.1 SWOT analysis..... 53
  - 4.2 Conclusion..... 55

Figure 4: Table of contents of the PDF target dossier. It is divided into four main sections: Target information, Disease information, Competitive landscape, and a final section for the conclusions.

presentation when instructed by experts on where to find the information. In response to a user’s question, the model automatically generates both the presentation and the PDF. LLMs are capable of generating text, but this is usually plagued by hallucinations, incomplete, and incorrect information. Here we show that an LLM can correct these deficiencies when instructed where to find the information (article repository, biological databases, code to run). The automatic generation of the target dossier can be seen as the first step of a more complex system. An automatic target dossier is an important step in standardizing the process. The automatic target dossier can be a valuable tool in helping different stakeholders in a pharmaceutical company, saving time and helping them to keep up to date quickly. Although the input of domain experts (biologists, physicians, chemists) in the creation of a target dossier remains key, this system allows a presentation and report to be generated automatically. This can be seen as a first step in creating different AI assistants that help the experts’ daily work. Different works show how AI is not meant to replace but to support them in daily tasks the experts, increasing their productivity and reducing their workload [42, 23]. Another improvement could be the utilization of a multimodal AI model. We used here an LLM for the reasoning steps, but our system retrieved and plotted images. A multimodal model can conduct reasoning by taking into account additional modalities.

## Target characteristics

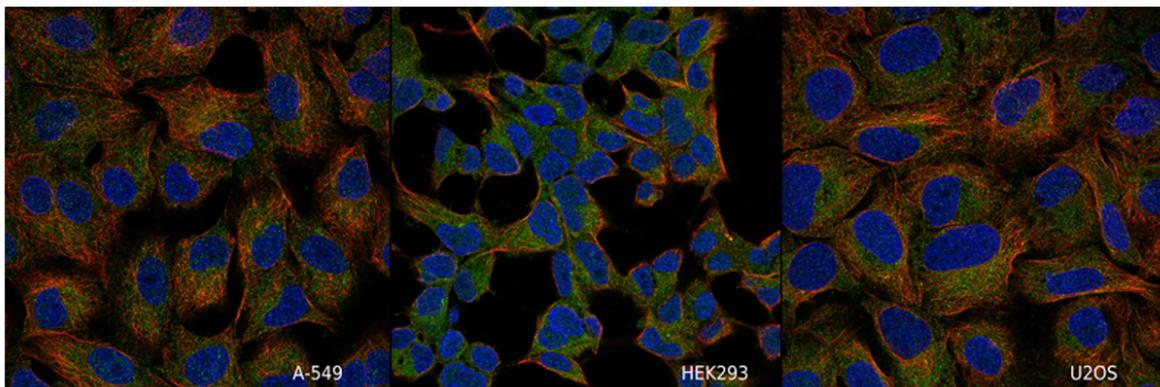
Target characteristics	
Similarity with monkeys	97.01%
Similarity with mice	98.94%
Similarity with rabbits	98.94%
Similarity with dogs	98.94%
Similarity with Guinea pigs	100.0%
Protein function	Ras proteins regulate cell proliferation by binding GDP/GTP and promoting oncogenic events through ZNF304-dependent silencing of tumor suppressor genes.

Source: UniProt and BLAST



Figure 5: Slide showing some characteristics of the target formatted as a table. The information is collected using UniProt and BLAST.

## Subcellular location



- Localized to the cytosol.
- Antigen: HPA072761



Source: Human Protein Atlas



Figure 6: Slide showing information about the subcellular location of the target available in the Human Protein Atlas.

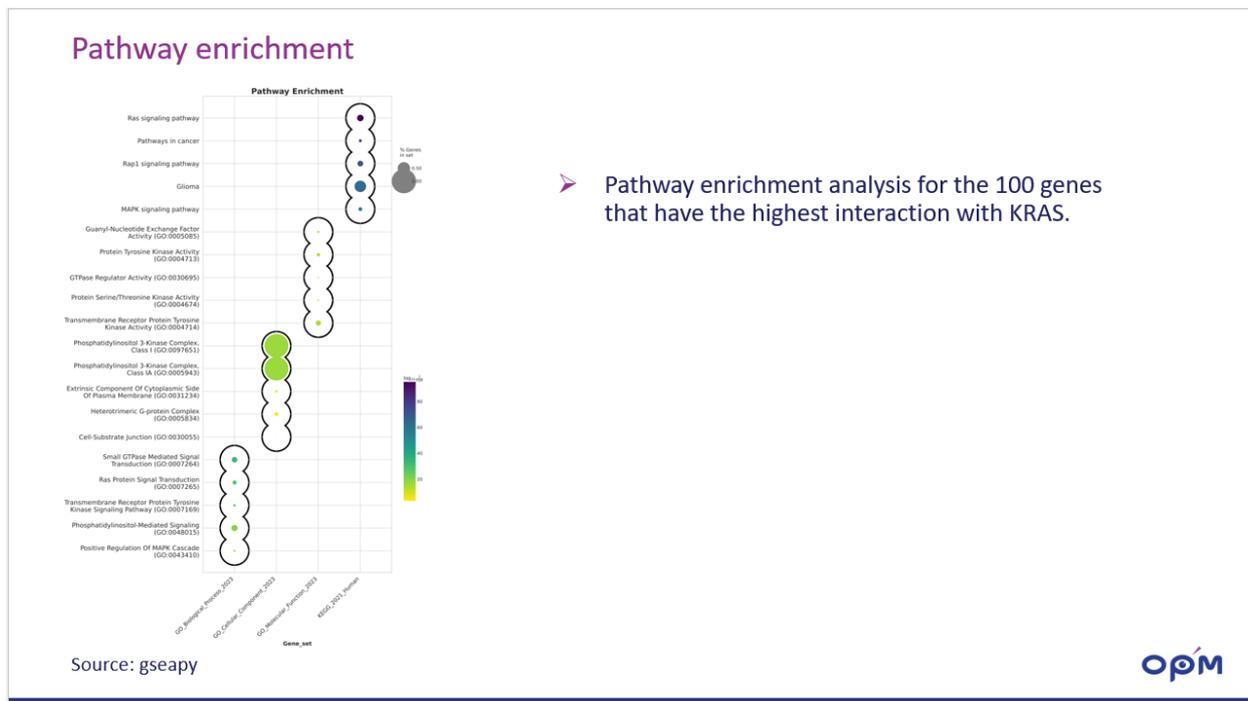


Figure 7: Slide showing the results of the pathway enrichment analysis. The list of genes which interact with the target is retrieved from the STRING database while the enrichment analysis is performed using the Python package GSEAPy.

## Acknowledgements

We would like to thank the members of the IT team, Jonathan Schmiedt and Thomas Wursten for their technical assistance during this project. We would also like to thank Maria Eugenia Riveiro, Kenji Shoji, and Oleksandr Levenets, Jan Hoflack for bringing their invaluable expertise and thorough insights.

## References

- [1] Steven M Paul et al. “How to improve R&D productivity: the pharmaceutical industry’s grand challenge”. In: *Nature reviews Drug discovery* 9.3 (2010), pp. 203–214.
- [2] S Eser et al. “Oncogenic KRAS signalling in pancreatic cancer”. In: *British journal of cancer* 111.5 (2014), pp. 817–822.
- [3] Barbara Bournet et al. “Targeting KRAS for diagnosis, prognosis, and treatment of pancreatic cancer: Hopes and realities”. In: *European journal of cancer* 54 (2016), pp. 75–83.
- [4] Joseph A DiMasi, Henry G Grabowski, and Ronald W Hansen. “Innovation in the pharmaceutical industry: new estimates of R&D costs”. In: *Journal of health economics* 47 (2016), pp. 20–33.
- [5] Richard K Harrison. “Phase II and phase III failures: 2013–2015”. In: *Nat Rev Drug Discov* 15.12 (2016), pp. 817–818.
- [6] Theodora Katsila et al. “Computational approaches in target identification and drug discovery”. In: *Computational and structural biotechnology journal* 14 (2016), pp. 177–184.
- [7] Karen M Mann et al. “KRAS-related proteins in pancreatic cancer”. In: *Pharmacology & therapeutics* 168 (2016), pp. 29–42.
- [8] Tohru Takebe, Ryoka Imai, and Shunsuke Ono. “The current status of drug discovery and development as originated in United States academia: the influence of industrial and academic collaboration on drug discovery and development”. In: *Clinical and translational science* 11.6 (2018), pp. 597–606.
- [9] Helen Dowden and Jamie Munro. “Trends in clinical success rates and therapeutic focus”. In: *Nat Rev Drug Discov* 18.7 (2019), pp. 495–496.
- [10] Kit-Kay Mak and Mallikarjuna Rao Pichika. “Artificial intelligence in drug development: present status and future prospects”. In: *Drug discovery today* 24.3 (2019), pp. 773–780.
- [11] W Patrick Walters and Regina Barzilay. “Applications of deep learning in molecule generation and molecular property prediction”. In: *Accounts of chemical research* 54.2 (2020), pp. 263–270.
- [12] Thomas Wolf et al. *HuggingFace’s Transformers: State-of-the-art Natural Language Processing*. 2020. arXiv: 1910.03771. URL: <https://arxiv.org/abs/1910.03771>.
- [13] Patrick Lewis et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. arXiv: 005.11401.
- [14] Jie Cai et al. “Advances in the epidemiology of pancreatic cancer: Trends, risk factors, screening, and prognosis”. In: *Cancer letters* 520 (2021), pp. 1–11.
- [15] Ji Luo. “KRAS mutation in pancreatic cancer”. In: *Seminars in oncology*. Vol. 48. 1. Elsevier. 2021, pp. 10–18.
- [16] Harrison Chase. *LangChain*. Oct. 2022. URL: <https://github.com/langchain-ai/langchain>.
- [17] Jianyuan Deng et al. “Artificial intelligence in drug discovery: applications and techniques”. In: *Briefings in Bioinformatics* 23.1 (2022), bbab430.
- [18] Madura KP Jayatunga et al. “AI in small-molecule drug discovery: a coming wave”. In: *Nat. Rev. Drug Discov* 21.3 (2022), pp. 175–176.
- [19] Duxin Sun et al. “Why 90% of clinical drug development fails and how to improve it?” In: *Acta Pharmaceutica Sinica B* 12.7 (2022), pp. 3049–3062.
- [20] Yujie You et al. “Artificial intelligence in cancer target identification and drug discovery”. In: *Signal Transduction and Targeted Therapy* 7.1 (2022), p. 156.
- [21] Albert Q. Jiang et al. *Mistral 7B*. 2023. arXiv: 2310.06825.
- [22] Jinze Bai et al. *Qwen Technical Report*. 2023. arXiv: 2309.16609.

- [23] Erik Brynjolfsson, Danielle Li, and Lindsey Raymond. *Generative AI at Work*. 2023. arXiv: 2304.11771 [econ.GM]. URL: <https://arxiv.org/abs/2304.11771>.
- [24] Yupeng Chang et al. *A Survey on Evaluation of Large Language Models*. 2023. arXiv: 2307.03109 [cs.CL]. URL: <https://arxiv.org/abs/2307.03109>.
- [25] J Clusmann et al. “The future landscape of large language models in medicine.” In: *Communications Medicine* 3.1 (2023), pp. 141–141.
- [26] Christopher J Halbrook et al. “Pancreatic cancer: Advances and challenges”. In: *Cell* 186.8 (2023), pp. 1729–1754.
- [27] Jason Hoelscher-Obermaier et al. *Detecting Edit Failures In Large Language Models: An Improved Specificity Benchmark*. 2023. arXiv: 2305.17553 [cs.CL]. URL: <https://arxiv.org/abs/2305.17553>.
- [28] Ayesha Juhi et al. “The capability of ChatGPT in predicting and explaining common drug-drug interactions”. In: *Cureus* 15.3 (2023).
- [29] Fenglin Liu et al. “A medical multimodal large language model for future pandemics”. In: *NPJ Digital Medicine* 6.1 (2023), p. 226.
- [30] Marissa Mock et al. “AI can help to speed up drug discovery—but only if we give it the right data”. In: *Nature* 621.7979 (2023), pp. 467–470.
- [31] Yujia Qin et al. *ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs*. 2023. arXiv: 2307.16789 [cs.AI]. URL: <https://arxiv.org/abs/2307.16789>.
- [32] Vipula Rawte, Amit Sheth, and Amitava Das. *A Survey of Hallucination in Large Foundation Models*. 2023. arXiv: 2309.05922 [cs.AI]. URL: <https://arxiv.org/abs/2309.05922>.
- [33] Yongliang Shen et al. *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face*. 2023. arXiv: 2303.17580 [cs.CL]. URL: <https://arxiv.org/abs/2303.17580>.
- [34] Karan Singhal et al. “Large language models encode clinical knowledge”. In: *Nature* 620.7972 (2023), pp. 172–180.
- [35] Xiangru Tang et al. “Medagents: Large language models as collaborators for zero-shot medical reasoning”. In: *arXiv preprint arXiv:2311.10537* (2023).
- [36] Shitao Xiao et al. *C-Pack: Packaged Resources To Advance General Chinese Embedding*. 2023. arXiv: 2309.07597 [cs.CL].
- [37] Yunzhi Yao et al. *Editing Large Language Models: Problems, Methods, and Opportunities*. 2023. arXiv: 2305.13172 [cs.CL]. URL: <https://arxiv.org/abs/2305.13172>.
- [38] Geyan Ye et al. *DrugAssist: A Large Language Model for Molecule Optimization*. 2023. arXiv: 2401.10334 [q-bio.QM]. URL: <https://arxiv.org/abs/2401.10334>.
- [39] Wayne Xin Zhao et al. *A Survey of Large Language Models*. 2023. arXiv: 2303.18223 [cs.CL]. URL: <https://arxiv.org/abs/2303.18223>.
- [40] Angels Balaguer et al. *RAG vs Fine-tuning: Pipelines, Tradeoffs, and a Case Study on Agriculture*. 2024. arXiv: 2401.08406 [cs.CL]. URL: <https://arxiv.org/abs/2401.08406>.
- [41] Elijah Berberette, Jack Hutchins, and Amir Sadovnik. *Redefining “Hallucination” in LLMs: Towards a psychology-informed framework for mitigating misinformation*. 2024. arXiv: 2402.01769 [cs.CL]. URL: <https://arxiv.org/abs/2402.01769>.
- [42] Kevin Zheyuan Cui et al. “The Productivity Effects of Generative AI: Evidence from a Field Experiment with GitHub Copilot”. In: *SSRN* (2024).
- [43] Yunfan Gao et al. *Retrieval-Augmented Generation for Large Language Models: A Survey*. 2024. arXiv: 2312.10997 [cs.CL]. URL: <https://arxiv.org/abs/2312.10997>.
- [44] Zorik Gekhman et al. *Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?* 2024. arXiv: 2405.05904 [cs.CL]. URL: <https://arxiv.org/abs/2405.05904>.
- [45] Qiao Jin et al. “GeneGPT: augmenting large language models with domain tools for improved access to biomedical information”. In: *Bioinformatics* 40.2 (Feb. 2024). Ed. by Jonathan Wren. ISSN: 1367-4811. doi: 10.1093/bioinformatics/btae075. URL: <http://dx.doi.org/10.1093/bioinformatics/btae075>.
- [46] Haoqiang Kang, Juntong Ni, and Huaxiu Yao. *Ever: Mitigating Hallucination in Large Language Models through Real-Time Verification and Rectification*. 2024. arXiv: 2311.09114 [cs.CL]. URL: <https://arxiv.org/abs/2311.09114>.
- [47] Jon M. Laurent et al. *LAB-Bench: Measuring Capabilities of Language Models for Biology Research*. 2024. arXiv: 2407.10362 [cs.AI]. URL: <https://arxiv.org/abs/2407.10362>.
- [48] Shervin Minaee et al. *Large Language Models: A Survey*. 2024. arXiv: 2402.06196 [cs.CL]. URL: <https://arxiv.org/abs/2402.06196>.
- [49] Oded Ovadia et al. *Fine-Tuning or Retrieval? Comparing Knowledge Injection in LLMs*. 2024. arXiv: 2312.05934.
- [50] Mayk Caldas Ramos, Christopher J. Collison, and Andrew D. White. *A Review of Large Language Models and Autonomous Agents in Chemistry*. 2024. arXiv: 2407.01603 [cs.LG]. URL: <https://arxiv.org/abs/2407.01603>.
- [51] S. M Towhidul Islam Tonmoy et al. *A Comprehensive Survey of Hallucination Mitigation Techniques in Large Language Models*. 2024. arXiv: 2401.01313 [cs.CL]. URL: <https://arxiv.org/abs/2401.01313>.
- [52] Jinge Wang et al. “Bioinformatics and biomedical informatics with ChatGPT: Year one review”. In: *Quantitative Biology* (2024).
- [53] Lei Wang et al. “A survey on large language model based autonomous agents”. In: *Frontiers of Computer Science* 18.6 (Mar. 2024). ISSN: 2095-2236. doi: 10.1007/s11704-024-40231-1. URL: <http://dx.doi.org/10.1007/s11704-024-40231-1>.
- [54] Hongjian Zhou et al. *A Survey of Large Language Models in Medicine: Progress, Application, and Challenge*. 2024. arXiv: 2311.05112 [cs.CL]. URL: <https://arxiv.org/abs/2311.05112>.
- [55] ChromaDB. *ChromaDB*. URL: <https://github.com/chroma-core/chroma>.

# **Annex B**

## **Integrating Large Language Models for Genetic Variant Classification**

---

# INTEGRATING LARGE LANGUAGE MODELS FOR GENETIC VARIANT CLASSIFICATION

---

A PREPRINT

Youssef Boulaimen<sup>1</sup>, Gabriele Fossi<sup>1</sup>, Leila Outemzabet<sup>1</sup>, Nathalie Jeanray<sup>1</sup>, Oleksandr Levenets<sup>1</sup>, Stéphane Gerart<sup>1</sup>, Sébastien Vachenc<sup>1</sup>, Salvatore Raieli<sup>1,2</sup>, and Joanna Giemza<sup>1</sup>

<sup>1</sup>Oncodesign Precision Medicine, 18 rue Jean Mazen, 21000 Dijon, France

<sup>2</sup>Corresponding author: sraieli@oncodesign.com

November 11, 2024

## ABSTRACT

The classification of genetic variants, particularly Variants of Uncertain Significance (VUS), poses a significant challenge in clinical genetics and precision medicine. Large Language Models (LLMs) have emerged as transformative tools in this realm. These models can uncover intricate patterns and predictive insights that traditional methods might miss, thus enhancing the predictive accuracy of genetic variant pathogenicity. This study investigates the integration of state-of-the-art LLMs, including GPN-MSA, ESM1b, and AlphaMissense, which leverage DNA and protein sequence data alongside structural insights to form a comprehensive analytical framework for variant classification. Our approach evaluates these integrated models using the well-annotated ProteinGym and ClinVar datasets, setting new benchmarks in classification performance. The models were rigorously tested on a set of challenging variants, demonstrating substantial improvements over existing state-of-the-art tools, especially in handling ambiguous and clinically uncertain variants. The results of this research underline the efficacy of combining multiple modeling approaches to significantly refine the accuracy and reliability of genetic variant classification systems. These findings support the deployment of these advanced computational models in clinical environments, where they can significantly enhance the diagnostic processes for genetic disorders, ultimately pushing the boundaries of personalized medicine by offering more detailed and actionable genetic insights.

**Keywords** Variants of Unknown Significance · Genomics · Deep Learning · Large Language Models

## 1 Introduction

The emergence of Next Generation Sequencing (NGS) (reviewed in [21]) has transformed the realm of genomics, enabling the sequencing of millions of DNA fragments. However, the interpretation of the NGS results poses significant challenges as the vast majority of identified variants are of unknown significance (VUS)[9, 8, 7]. An accurate prediction of such variants can pave the way to a better understanding of disease mechanisms, enabling personalized medicine and the discovery of new therapeutic targets.

Over the years, many computational tools and datasets have been developed to help predict the effects of variants. Early tools like PolyPhen and SIFT used sequence homology and protein structure information to predict the impact of missense mutations [2, 19]. Other models, such as CADD [14], combine multiple annotations into a single score to indicate variant pathogenicity.

The promising results of Large Language Models (LLMs) in Natural Language Processing (NLP) tasks have led to their adaptations in the fields of genomics and proteomics. LLMs are complex models that use the Transformer architecture[24]. One particular component of the Transformer architecture is self-attention, which enables the model to weigh the importance of different parts of the input data dynamically. This mechanism allows the models to consider the entire sequence context, making it particularly effective in handling long-range dependencies and interactions within

the data. A remarkable example of the high potential of LLMs in proteomics is ESMFold, a protein language model that can predict protein structures using protein sequences[18].

In variant effect prediction (VEP), exploiting the capabilities of self-attention can be beneficial as it allows the model to account for not only specific mutations but also the entire genetic background and associated protein sequences, providing a comprehensive view of the molecular context. LLMs such as GPN-MSA, ESM1b, and Alphamissense have shown promise in predicting variant pathogenicity. GPN-MSA is a DNA language model trained on MSA (Multiple Sequence Alignment) of 100 species which leverages evolutionary information in predicting pathogenicity scores for all possible nucleotide substitutions in the genome[3]. ESM1b is a protein language model that predicts the pathogenicity for all 20 possible amino acids, without relying on homology and taking into account all protein isoforms[4]. As for Alphamissense, it is first trained to predict protein structures from the protein sequences and later fine-tuned for pathogenicity prediction[7]. These models are considered as state of the art in VEP.

We hypothesize that integrating these models may offer significant advantages. By combining their predictions, we can not only capitalize on their strengths and address their weaknesses but also provide a more comprehensive prediction leveraging both DNA and protein data. We adopted this integrative approach, using machine learning models for the combination of the scores, in order to develop a more accurate and comprehensive tool for predicting variant pathogenicity.

## 2 Materials & Methods

### 2.1 Data

In this study, we utilized the ProteinGym dataset[20], a comprehensive resource developed to facilitate the evaluation of mutation effect predictors. This dataset is divided into two primary benchmarks: substitution benchmark and indel benchmark. For our analysis, we focused exclusively on the substitution benchmark of the ProteinGym dataset (accessed on 3/22/24), which includes approximately 2.7 million missense variants characterized across 217 Deep Mutational Scanning (DMS) assays[11] and encompasses data on 2,525 clinical variants. We examined specifically two distinct segments within this benchmark: the clinical variants substitutions dataset and the raw substitutions dataset. The raw substitution dataset is extensive and contains 61 columns. Crucial to our study are the columns indicating the chromosome and the exact genomic location of each variant, as well as reference and alternative alleles, which detail the nucleotide changes. Also integral to our analysis are the columns detailing the corresponding protein position and the amino acid changes. These are linked via the transcript ID, which connects the genomic data to specific protein transcripts, thereby facilitating cross-references between the genetic and protein data. Moreover, the dataset includes columns that categorize the clinical significance of each variant, classifying them as either pathogenic or benign. The clinical substitutions dataset contains the transcript ID to ensure consistent referencing across the datasets. It records both the position within the protein and the reference and alternate amino acids involved in each substitution. Additionally, it provides the sequences before and after mutations, along with the DMS\_bin\_score that classifies each protein substitution as benign or pathogenic.

GPN-MSA’s HuggingFace repository (accessed on 3/12/24) provides predictions for all possible SNPs in the human genome. Using the chromosomes and positions of the variants from the raw substitutions dataset, we queried the scores for all three possible nucleotide substitutions using Tabix[17]. The lower the score of GPN, the more pathogenic the variant.

We employed the ProteinGym substitution dataset to compute the ESM1b scores, which include reference protein sequences along with detailed mutation information such as positions and the specific amino acids involved. The ESM1b code was sourced from its GitHub repository. This model takes protein sequences as input and produces scores for all 20 possible amino acid substitutions at each position within the sequence, which can lead to extensive output files and considerable processing times. To streamline this process, we modified the ESM1b code to focus on scoring only the specified mutation positions from the dataset. This targeted approach significantly reduced both the output file size and the processing time. The lower the score of ESM1b, the more pathogenic the variant. A Log Likelihood Ratio (LLR) threshold of -7.5 was used to distinguish between pathogenic and benign variants[4].

For our analysis, we also utilized predictions from AlphaMissense, accessible through the file AlphaMissense-aa-substitutions.tsv.gz (4/8/2024). This dataset contains predictions for all conceivable single amino acid substitutions within 20,000 UniProt canonical isoforms, totaling approximately 216 million protein variants. Integrating AlphaMissense predictions posed significant challenges due to discrepancies in protein identifier systems. AlphaMissense uses UniProt accession numbers, whereas the ProteinGym dataset relies on NCBI’s RefSeq protein IDs. To address this, we utilized the UniProt ID mapping tool to align the datasets, successfully mapping 2,415 out of the 2,525 proteins from

ProteinGym. This mapping allowed us to accurately link UniProt accession numbers to the corresponding mutations in ProteinGym and retrieve the necessary AlphaMissense scores for our analysis.

## 2.2 Data Processing

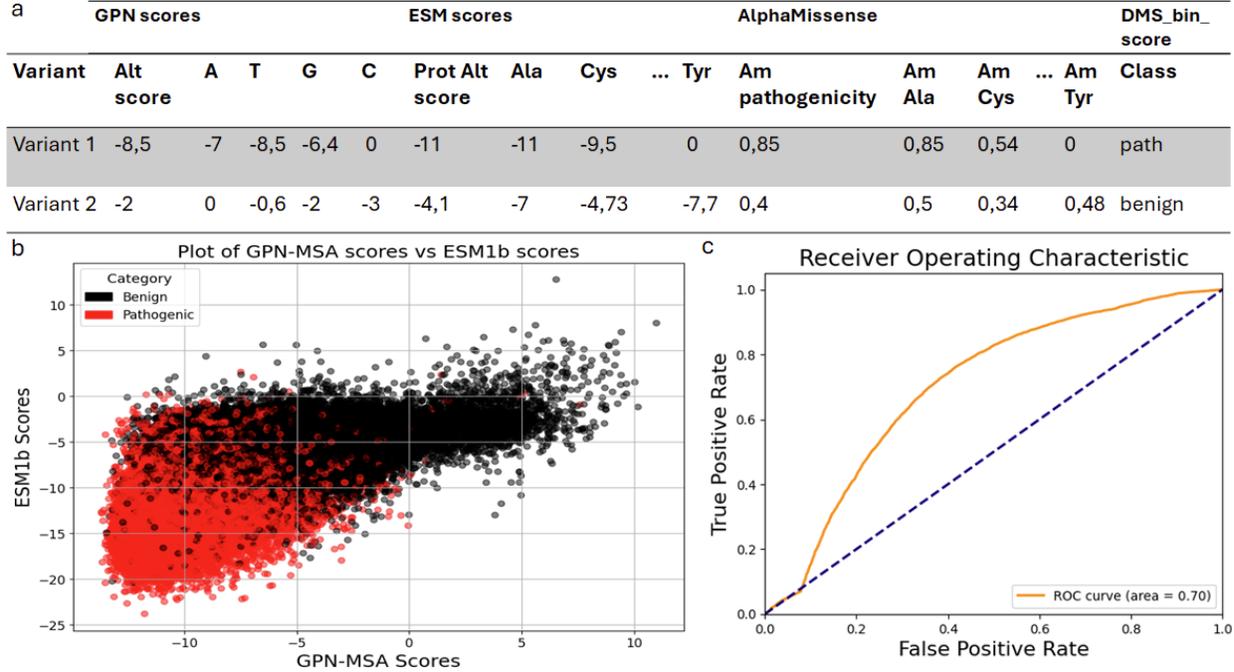


Figure 1: Dataset presentation: **a. Dataset Samples:** This table provides a representative sample of the dataset utilized for this study, showcasing both observed and potential mutation scores derived from three distinct models: GPN scores, ESM scores, and AlphaMissense scores. Each entry represents the score assigned by the respective model to various attributes such as nucleotide changes (A, T, G, C) and amino acid substitutions (e.g., Ala, Cys, Tyr), as well as the observed scores in columns Alt\_score, Prot\_Alt\_score, and Am\_pathogenicity. Nucleotides and Amino Acids with a score of value 0 correspond to the reference allele or protein. The "DMS\_bin\_score" column indicates the clinical classification of the mutation as either "pathogenic" or "benign." **b. Dataset Visualization:** This plot represents the distribution of the variants' observed scores of GPN-MSA as Alt Score on the X-axis and ESM1b as Prot Alt Score on the Y-axis. Red data points represent the variants classified as pathogenic by DMS\_bin\_score whereas the black data points are classified as benign. **c. Optimal Threshold for GPN:** This ROC curve illustrates the performance of the GPN model in discriminating between pathogenic and benign genetic variants. The x-axis represents the False Positive Rate (FPR), and the y-axis represents the True Positive Rate (TPR), across various threshold levels. The orange line depicts the actual ROC curve, which shows how the TPR and FPR change with different thresholds. The area under the curve (AUC) is 0.70, indicating the model's overall ability to distinguish between the classes; a value of 1.0 represents a perfect classifier, and a value of 0.5 represents a random guess. The dashed blue line represents the line of no discrimination, which serves as a baseline comparison. The optimal threshold for classification is found by maximizing the difference between TPR and FPR.

This analysis differentiated between two types of scores. Firstly, the observed mutation scores, are calculated for mutations that are actually present in the dataset and represent clinically observed mutations in genomic or protein sequences. These scores provide direct insights into the impact of a specific, known mutation for one alternative nucleotide or amino acid (Fig.1.a columns Alt score, Prot Alt score, and AM pathogenicity). These observed mutations have an experimental annotation in the DMS\_bin\_score column. Secondly, the potential mutation scores, which speculate on the theoretical impact of all conceivable mutations at each position within the genome or protein sequence. For GPN-MSA, this involves generating four potential scores corresponding to the four possible nucleotide changes at each genomic position. In the case of ESM1b and AlphaMissense, scores are generated for each of the 20 possible amino acid substitutions at each position in the protein sequence ((Fig.1.a columns A, T, Ala, Am Ala...)). In cases where the reference and alternative alleles or amino acids are identical, a score of zero is assigned, reflecting no change

or impact due to the mutation. The potential mutations do not have a pathogenicity classification, except for the one corresponding to the observed mutation.

The first step of the data processing was merging the GPN-MSA and ESM1b scores using the transcript ID and protein information from both ProteinGym datasets, thus obtaining a dataset of 59,593 rows. This dataset was used for the preparation of the training and testing sets for the deep learning models. The splitting was performed in a manner that keeps the most ambiguous and difficult-to-predict data points in the test set by selecting a threshold for both scores. After visualizing the distribution (Fig.1), we selected the variants with scores between -5 to -10.

AlphaMissense scores were added later, resulting in a small reduction of the dataset to 49,554 rows due to the proteins lost during the mapping. A new test set was generated by merging the previous test set with the new dataset containing the 3 scores. Using the threshold resulted in a test set of 16165 rows and a training set of 33,389 rows with balanced proportions of 16,588 pathogenic and 16,801 benign variants.

The last step was to assign a threshold for the GPN-MSA scores to enable the comparison between all models. To find the optimal threshold, one approach is to maximize the difference between True Positive Rate (TPR) and False Positive Rate (FPR). This is typically done by calculating TPR and FPR for each possible threshold using the Receiver Operating Characteristic (ROC) curve, and then identifying the point where the difference between TPR and FPR is the greatest. This index corresponds to the optimal threshold from the evaluated thresholds array. For the GPN-MSA scores, this method was used to identify a threshold that optimally differentiates between pathogenic and benign variants, ensuring accurate and clinically relevant comparisons across all models. The threshold found for the GPN-MSA scores was -7 with an optimal FPR of 0.41 and optimal TPR of 0.759. For the other models, AlphaMissense predictions were taken directly from its `am_class` output, which labels variants as either Pathogenic, Benign or Ambiguous. For ESM1b, variants with a score of -7.5 or below were considered pathogenic as described in the paper[4].

## 2.3 Model Architectures

Various machine learning models were used for the training, namely XGBoost (XGB)[6], Random Forest (RF)[5], and Neural Networks. All models were trained using different sets of pathogenicity scores as features and the `DMS_bin_score` as the target variable. The ensemble models such as XGB and RF were trained using default parameters, as the fine-tuning of such models requires using grid search, which demands extensive amounts of time for limited improvements. As for Neural Networks, several architectures were employed. As the protein and DNA models provide a different number of scores, we decided to explore both Multi-input Neural Networks that take each score separately and Single-input Neural Networks that take all of the scores altogether. The architectures and parameters were explored and the optimal values were selected through a process of trial and error. The architectures and parameters were systematically optimized through an iterative process of trial and error to determine the optimal configuration.

### 2.3.1 Single input Neural Networks

The model architecture included a single input layer to handle the scores. This input was passed through a Dense layer with 64 units and a LeakyReLU activation function. A Dropout layer with a rate of 0.5 followed to prevent overfitting. The dropout layer was connected to another Dense layer with 128 units and a LeakyReLU activation function. The final output layer was a single unit Dense layer with a sigmoid activation function for binary classification.

The model was compiled using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.001 and binary cross-entropy loss function. Training was performed for up to 350 epochs with a batch size of 32. Early Stopping with a patience of 10 epochs and ReduceLROnPlateau with a factor of 0.2 and threshold of 0.0001 were employed to prevent overfitting and adjust the learning rate, respectively[1].

### 2.3.2 Multi-input Neural Networks

The model included three input layers to handle the different scores. Each branch began with a Dense layer of 64 units, followed by Batch Normalization and ReLU activation. The outputs from these three branches were concatenated into a single tensor. This concatenated tensor was then passed through two Dense layers with 256 and 128 units, respectively. Each dense layer was followed by Batch Normalization, ReLU activation, and Dropout with a rate of 0.5 to prevent overfitting. The final output layer was a single unit Dense layer with a sigmoid activation function, appropriate for binary classification tasks. The model was compiled using the Stochastic Gradient Descent (SGD) optimizer with a learning rate of 0.01 and the binary cross-entropy loss function. Training was conducted for up to 350 epochs with a batch size of 32. To prevent overfitting and adjust the learning rate during training, Early Stopping with a patience of 50 epochs and ReduceLROnPlateau with a factor of 0.2 and threshold of 0.0001 were utilized.

## 2.4 Case study methodology

Case studies were conducted on select variants to confirm the model’s prediction and showcase its efficiency as described in 3.4. We followed a specific methodology in order to extract the necessary data about these variants. First, we look into the protein information in Uniprot using the UniProt ID in our dataset. We first note the general data about the protein such as the function and the structure. Next, we download the PDB file for the AlphaFold structure of the protein from AlphaFold’s website. The PDB file is then used to visualize the protein structure with Pymol. Pymol enables to change the specific residue in a specific position to another residue and observe the changes in the protein structure using the Mutagenesis panel. We use this to obtain the structure for the protein with the mutation we need. This way it is possible to visualize both the wild-type protein’s structure as well as the mutated protein. Later we explore the interactions of the WT and mutated residues with the neighbouring structures in a radius of 3.5 Angstroms. This enables us to hypothesize on the possible outcomes of the mutation. Next, we investigate the Disease & Variants section which contains the known diseases the protein is involved in and the variants that are possibly implicated. For the well-annotated and well-studied variants, links for scientific papers are provided which are also investigated. In case the mutation we investigate is not included in the Disease & Variants section, the variant viewer panel in UniProt. This panel contains information on several mutations in the protein such as the effect of the mutation from different databases but also potential diseases in which the mutation can be involved. The main databases used are usually ClinVar, gnomAD, and dbSNP. Finally, we explore the Family & Domains section to check whether the mutation is involved in a functional domain of the protein. This section also provides scientific papers of studies conducted on specific regions of the proteins.

## 2.5 Technical Details

The computational analysis was conducted using Python 3.8 in a Jupyter Notebook environment. The platform used was a Linux operating system (Linux-5.15.0-1044-nvidia-x86\_64-with-glibc2.27). The hardware specifications include a 64-bit processor architecture (x86\_64) with 40 physical cores, providing substantial computational power for parallel processing tasks. The system was equipped with 503 GB of total memory (RAM). The total disk space available was 438 GB. A key feature of the setup is the presence of eight NVIDIA Tesla V100-SXM2-32GB GPUs, each with 32 GB of dedicated memory. The CUDA version employed was 12.3, supported by NVIDIA driver version 545.23.08.

## 3 Results

The evaluation of several machine learning models was conducted to assess their effectiveness in predicting the pathogenicity of genetic variants. In this section, we will describe a correlation analysis between the scores of GPN-MSA, ESM1b, and AlphaMissense, benchmarking of the various models and a selection of the optimal set of features. Next, we will compare our best model’s performance to state-of-the-art tools to contextualize its performance. Finally, we performed case studies to further showcase our model’s utility in real-world applications

We utilized a dataset composed of various genetic scores derived from GPN-MSA, ESM1b, and AlphaMissense. These scores were the primary inputs for the predictive models. Four different feature sets were considered as inputs for the predictive models:

- Potential scores from GPN-MSA and ESM1b: This feature set contains all four possible predictions from GPN-MSA and twenty predictions from ESM1b. It will test the impact of integrating both DNA and protein data and will also serve as a basis for comparison with other models that will incorporate Alphamissense scores.
- Potential scores from GPN-MSA, ESM1b, and AlphaMissense: This feature set contains all four possible predictions from GPN-MSA and twenty predictions from ESM1b and Alphamissense. This feature set will assess the utility of adding Alphamissense scores to the model.
- Observed Scores from GPN-MSA, ESM1b, and AlphaMissense: This feature set contains only one score for each model. This score corresponds to the score of the observed mutation in the dataset. Using this feature set we will test the necessity of using all the potential scores.
- Observed and Potential Scores from GPN-MSA, ESM1b, and AlphaMissense: This feature set includes the scores for observed mutations for each model alongside all the possible model predictions. Observed scores represent the actual clinical mutations identified in the dataset, reflecting real-world genetic variations with known clinical significance. Potential scores, on the other hand, encompass a broad range of hypothetical mutations that provide a comprehensive view of possible genetic variations. Including both sets of scores in the feature set introduces a duplication of the observed scores, effectively weighting them more heavily in the

model. This feature engineering approach is designed to emphasize the real, clinically validated mutations captured by the observed scores, thereby potentially improving the model’s performance.

The models were trained on a subset of the data and evaluated on a separate test set specifically chosen to include the most ambiguous and challenging variants. This approach was used to evaluate the model performance as strictly as possible, ensuring that the models are robust and effective even under the most difficult conditions. The test set composition included a balanced mix of 7902 pathogenic and 8263 benign variants.

### 3.1 Correlation Analysis

The correlation analysis quantified the degree of linear association between the observed scores from GPN-MSA, ESM1b, and AlphaMissense using Pearson correlation coefficients. This analysis provides insights into the prediction trends among these models, evaluating the necessity for an integrated model approach. GPN-MSA and ESM1b show a positive correlation (0.6779), suggesting these models tend to align in their predictions. In contrast, GPN-MSA and AlphaMissense have a negative correlation (-0.7259), and ESM1b and AlphaMissense exhibit an even stronger negative correlation (-0.8104). The negative correlations between AlphaMissense and the other two models (GPN-MSA and ESM1b) are expected, given that AlphaMissense assigns higher scores to pathogenic variants while the other two models assign higher scores to benign variants. The stronger correlation between ESM1b and AlphaMissense is due to their focus on protein-level pathogenicity predictions, while GPN-MSA predicts at the DNA level, explaining its lower correlation with both models. Overall, the results justify the need for an integrated model, as the individual models capture different aspects of the data, providing a more comprehensive and nuanced analysis when combined.

	GPN-MSA	ESM1b	AM
GPN-MSA	1	0,6779	-0,7259
ESM1b	0,6779	1	-0,8104
AM	-0,7259	-0,8104	1

Figure 2: **Correlation Analysis** This table illustrates the correlation matrix between the observed scores from the GPN-MSA, ESM1b, and AlphaMissense models. The correlation coefficients quantify the degree to which these models agree or disagree on the pathogenic potential of the mutations, providing insight into their comparative analytical behaviors.

### 3.2 Model performances

Here we evaluate the performances of the models across the different feature sets. The model performances are detailed in Fig.3. Overall, the choice of the models didn’t have a significant impact on the results. However, the evaluation revealed varying performance across the different feature configurations. All models showed similar performance with the GPN+ESM feature set, with accuracies around 0.75 and ROC-AUC just below 0.83. The multi-input neural network and random forest slightly outperformed the XGBoost and single-input neural network in terms of ROC-AUC. The incorporation of AlphaMissense scores improved performance across all models, particularly the Random Forest model with a ROC-AUC of 0.872. The single-input neural network excelled with the Observed Scores configuration, achieving the highest accuracy 0.808 and ROC-AUC of 0.877. The feature set that included both Observed Scores and Potential Scores provided the best overall results, particularly for the multi-input neural network and random forest, both achieving high accuracies and ROC-AUC scores above 0.89.

Features	Model	Evaluation metrics				
		Accuracy	Precision	Recall	F1 Score	ROC-AUC
Potential scores (GPN+ESM)	Multi-input Neural Network	0.7558	0.742	<b>0.767</b>	0.754	<b>0.8297</b>
	Single input Neural Network	0.749	0.735	0.7599	0.747	0.82
	XGBoost	<b>0.758</b>	<b>0.746</b>	0.766	<b>0.7559</b>	0.8293
	Random Forest	0.7558	0.7446	0.7617	0.753	<b>0.8297</b>
Potential scores (GPN+ESM+ AlphaMissense)	Multi-input Neural Network	0.79511	0.7829	0.80359	0.793	0.86983
	Single input Neural Network	0.782	0.7681	0.7948	0.7812	0.858
	XGBoost	0.794	0.78	0.8048	0.7925	0.8684
	Random Forest	<b>0.798</b>	<b>0.784</b>	<b>0.81</b>	<b>0.797</b>	<b>0.872</b>
Observed scores (GPN+ESM+ AlphaMissense)	Single input Neural Network	<b>0.808</b>	<b>0.804</b>	0.802	0.803	<b>0.877</b>
	XGBoost	<b>0.808</b>	0.799	0.81	<b>0.805</b>	0.876
	Random Forest	0.805	0.799	<b>0.804</b>	0.801	0.871
Observed scores + Potential scores (GPN+ESM+ AlphaMissense)	Multi-input Neural Network	0.825	<b>0.8229</b>	0.8189	0.8209	<b>0.894</b>
	XGBoost	0.822	0.8126	0.826	0.819	0.891
	Random Forest	<b>0.829</b>	0.821	<b>0.832</b>	<b>0.826</b>	<b>0.894</b>

Figure 3: **Comparative performance of Machine Learning models in genetic variant classification:** This table presents the benchmarking results of different machine learning models using various combinations of features derived from GPN-MSA, ESM1b, and AlphaMissense. The models evaluated include multi-input neural networks, single-input neural networks, XGBoost, and Random Forest, each tested across four distinct feature sets: GPN+ESM potential scores, GPN+ESM+AlphaMissense potential scores, observed scores from GPN+ESM+AlphaMissense, and a combination of observed and potential scores from GPN+ESM+AlphaMissense.

### 3.3 Performance comparison state of the art models

Next, we will compare the Multi-input Neural Network model trained on Observed + Potential scores to state-of-the-art tools. This analysis aims to assess the real-world applicability of our approach, particularly in its ability to accurately predict Variants of Uncertain Significance (VUS). By benchmarking our model against existing leading tools, we aim to demonstrate its effectiveness and potential advantages in clinical and research settings. This comparison will help us understand how well our integrated model performs in practical scenarios, ensuring its utility in improving genetic variant classification.

#### 3.3.1 Performance comparison state of the art models: ProteinGym

In this study, the Multi-input NN model trained on Observed+Potential scores was evaluated alongside AlphaMissense, GPN-MSA, and ESM1b, using the test dataset of 16,165 genetic variants. This dataset was specifically chosen to include the most ambiguous and challenging variants to evaluate model performance as strictly as possible, ensuring robustness and effectiveness under difficult conditions. The models' performance was assessed by comparing their predictions against the experimental annotations provided in the ProteinGym's DMS\_bin\_score. For the assessment, AlphaMissense predictions were taken directly from its am\_class output, which labels variants as either Pathogenic, Benign or Ambiguous. For ESM1b, variants with a score of -7.5 or below were considered pathogenic. The threshold for GPN-MSA was set at -7, as detailed in the Materials and Methods. The initial comparison of the models was performed on all 16,165 variants of the test set (Fig.4a). This comparison aims to provide a baseline comparison of the models on the challenging variants of the test set. The analysis demonstrated that the integrated model outperformed the others, achieving an accuracy of 82.54%. AlphaMissense and ESM1b showed comparable performances with accuracy of 74.58% and 73.84% respectively, while GPN-MSA lagged at 67.03%. For the second analysis, variants classified as

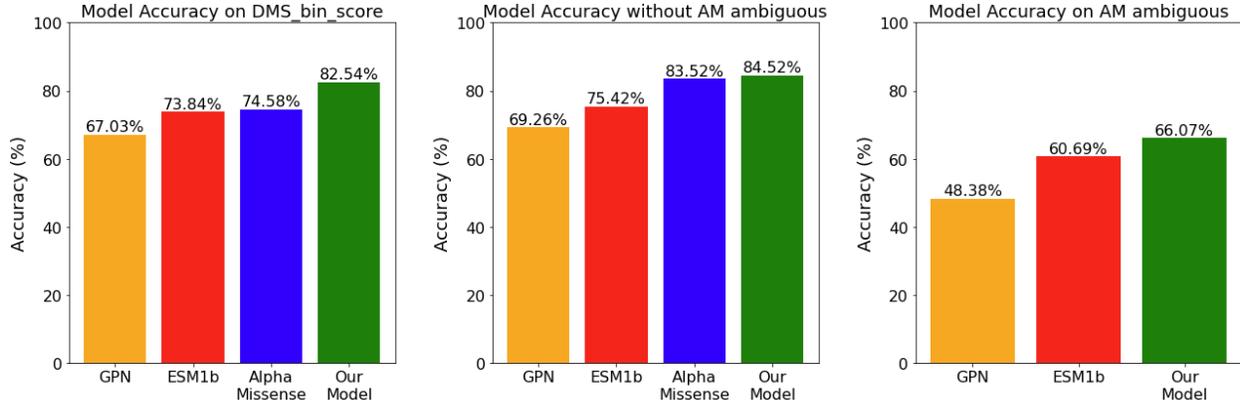


Figure 4: **Model Performance Across Different Conditions and Datasets:** (a): This panel displays the accuracy of individual models on the DMS\_bin\_score. The bars represent the accuracy of the AlphaMissense model, the integrated Model (combining GPN-MSA, ESM1b, and AlphaMissense), ESM1b, and GPN. Here, the integrated Model shows a strong performance with an accuracy of 82.54%, followed by AlphaMissense at 74.58%. ESM1b and GPN exhibit lower accuracies at 73.84% and 67.03% respectively. (b): This panel illustrates the model accuracy after removing variants classified as ambiguous by AlphaMissense. This graph provides insights into how the clarity of variant classification affects model performance. The integrated Model achieves the highest accuracy at 84.51%, followed by AlphaMissense at 83.51%, and ESM1b at 75.41%. GPN shows significantly lower accuracy at 69.26%, suggesting it is more affected by the removal of ambiguous variants compared to the other models. (c): This panel focuses specifically on the accuracy of models in predicting AlphaMissense classified ambiguous variants, highlighting the challenges in handling ambiguous genomic data. The integrated Model maintains the highest accuracy at 66.07%, demonstrating its robustness even in uncertain conditions. ESM1b and GPN show reduced accuracies at 60.69% and 48.38% respectively.

ambiguous by AlphaMissense were excluded from the analysis, reducing the test set to 14,435 variants (Fig.4b). This comparison aimed to provide a fairer assessment by considering only two predictions, removing the variants that are difficult for AlphaMissense to predict. All models displayed improved accuracy. The integrated model still outperformed AlphaMissense, achieving accuracies of 84.52% and 83.52% respectively. The last analysis focused on the 1,730 variants categorized as ambiguous by AlphaMissense (Fig.4c). Here we will measure how the models perform on a smaller subset of hard-to-predict variants according to AlphaMissense. In this challenging subset, the integrated model again showed robust performance, achieving the highest accuracy of 66.07%. ESM1b and GPN, presented accuracies of 60.69% and 48.38% respectively. It's important to note that we solely relied on accuracy for model evaluation in this study. AlphaMissense outputs three classifications (Pathogenic, Benign, Ambiguous), rendering the calculation of other common metrics like ROC-AUC more complex.

### 3.3.2 Performance comparison state of the art models: ClinVar

In addition to the DMS annotation, the ClinVar[15] classifications were added to the test dataset for further comparison. This merge implied the loss of a single row bringing the dataset to 16,164 variants. The ClinVar data provides an array of different classifications. Variants with labels 'Uncertain significance' or 'Conflicting classifications of pathogenicity' were classified as Ambiguous. The other variants were either labeled Benign or Pathogenic. Comparing the models' performances on the ClinVar dataset provides an additional layer of analysis on a clinically relevant dataset, offering valuable insights into the models' effectiveness in handling Variants of Uncertain Significance (VUS).

First, we performed the comparison of each model against the ClinVar classifications Fig.5a. This comparison was motivated by the need to evaluate the overall accuracy of each model. In this analysis, AlphaMissense should have an advantage as it can directly classify ambiguous variants, whereas other models, which only provide benign or pathogenic predictions, would generate false predictions for these ambiguous variants. Despite this, the integrated model still outperformed AlphaMissense in this task. The integrated model achieved an accuracy of 79.16%, AlphaMissense 72.07%, ESM1b 70.93%, and GPN 64.33%.

Next, we tested the models' performances on clearly classified variants by removing the 715 ambiguous variants of the ClinVar dataset Fig.5b. This analysis was conducted to provide a more straightforward comparison, focusing solely on benign and pathogenic classifications without the complexity introduced by ambiguous variants. This analysis shows an

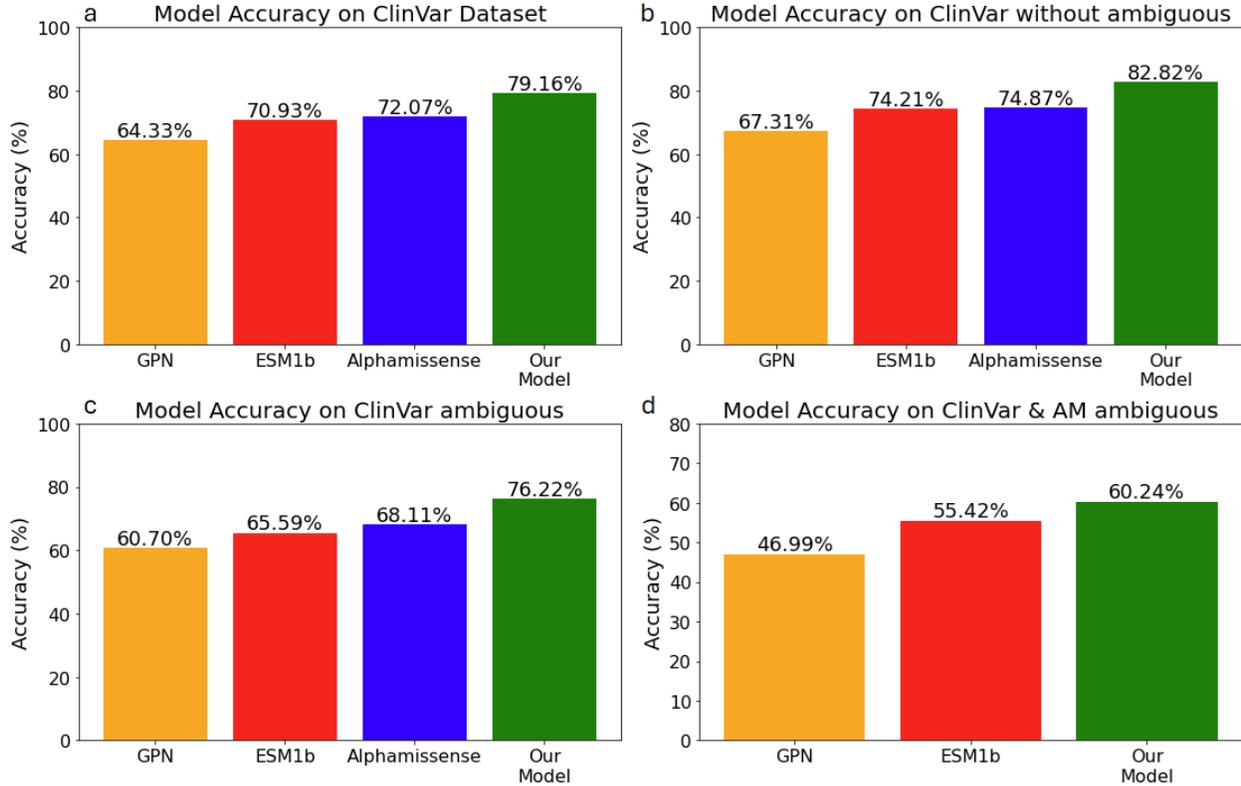


Figure 5: **Model Performance on ClinVar Dataset:** (a): This panel displays the accuracy of AlphaMissense, the integrated Model, ESM1b, and GPN when tested against the ClinVar dataset. The ClinVar dataset provides three different classes: Pathogenic, Benign, and Ambiguous. The integrated Model shows the highest accuracy at 79.16%, followed by AlphaMissense at 72.07%, ESM1b at 70.93%, and GPN showing the lowest accuracy at 64.33%. (b): This panel illustrates the accuracy of the same models on the ClinVar dataset after the removal of the 715 ambiguous variants. The performance of all models is generally improved. The integrated Model leads in accuracy at 82.82%, demonstrating its effectiveness in classifying clearly defined genetic variants. This is followed by AlphaMissense at 74.87%, ESM1b at 74.21%, and GPN at 67.31%. (c): This panel focuses on variants classified as ambiguous in the ClinVar dataset while using DMS as ground truth. The graph illustrates that the integrated Model maintains superior performance even in this subset, achieving an accuracy of 76.22%. AlphaMissense follows at 68.11%, ESM1b at 65.59%, and GPN shows the least accuracy at 60.70%. (d): This panel examines the accuracy of the models on a combined subset of variants classified as ambiguous by both ClinVar and AlphaMissense, with DMS used as ground truth for performance evaluation. The integrated Model continues to show superior performance in this challenging scenario with an accuracy of 60.24%. This is followed by ESM1b at 55.42%, and GPN at 46.99%, further validating the robustness of the integrated Model.

overall better accuracy across all the models compared to Fig.5.a with the integrated model outperforming the other models with an accuracy of 82.82%, followed by AlphaMissense at 74.87%, ESM1b at 74.21%, and GPN at 67.31%.

Subsequently, the performance of the models on the 715 ambiguous variants of ClinVar was assessed using the DMS\_bin\_score as the ground truth (Fig.5.c). This task aimed to evaluate how effectively each model discriminates between variants with uncertain significance. The results indicated weaker performances compared to the overall dataset (Fig.5.a). Nonetheless, the integrated model again outperformed the other models with an accuracy of 76.22%, showcasing its efficacy in classifying ambiguous variants. AlphaMissense, ESM1b, and GPN showed accuracies of 68.11%, 65.59%, and 60.70%, respectively.

Finally, the performance of the integrated model, ESM1b, and GPN-MSA was assessed on a small dataset of 83 variants classified as ambiguous by both AM and ClinVar (Fig.5d). This comparison was motivated by the need to evaluate the models on the most challenging subset, where both AlphaMissense and ClinVar classifications had flagged the variants as ambiguous. The DMS\_bin\_score was used as the ground truth for this evaluation. The integrated model continued

to display robust results with 60.24% correct predictions, followed by ESM1b with 55.42% and GPN with 43.37%. Assessing the models on this difficult subset underscores the integrated model’s capability to manage the complexities associated with ambiguous genetic data, further validating its robustness and reliability in real-world applications where uncertain classifications are prevalent.

### 3.4 Case studies

Here we investigate examples of variants classified as VUS by ClinVar and for which our model aligns with DMS\_bin\_score as opposed to the other three models. These case studies help showcase the utility of our model in real-life scenarios but also prove our model effectively learned underlying data from the state of the art and makes accurate predictions autonomously. We also looked into the cases where all the models’ predictions align with the DMS\_bin\_score except for our model and found no such cases.

#### 3.4.1 Case study of LZTR1

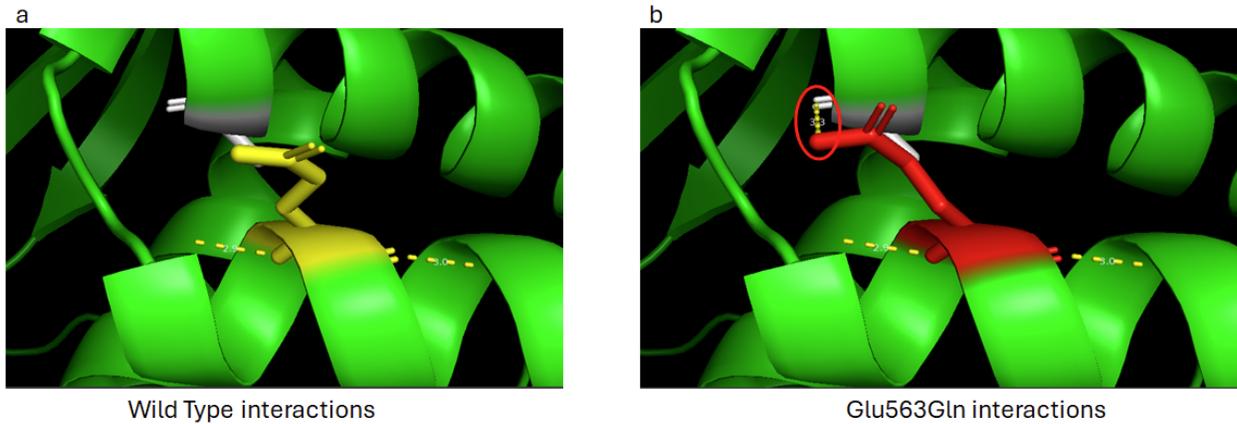


Figure 6: **Visualization of the E563Q mutation in the Leucine Zipper-like Transcriptional Regulator 1 (LZTR1) protein:** (a): Visualization of the Wild Type residue (Glu563) and its interactions. (b): Visualization of the mutated residue (Gln563) and its interactions. The Glu563Gln mutation causes a new interaction with an alpha helix (circled in red).

For the first case study, we focused on the E563Q mutation in the Leucine Zipper-like Transcriptional Regulator 1 (LZTR1). The Leucine Zipper-like Transcriptional Regulator 1 operates within the Golgi apparatus. It acts as a negative regulator of RAS-MAPK signaling by controlling Ras levels and decreasing Ras association with membranes. LZTR1 is also hypothesized to interact with the CUL3 ubiquitin ligase complex, which facilitates the degradation of redundant proteins. Functionally, LZTR1 is believed to act as a tumor suppressor.

Our model classified the E563Q mutation as pathogenic, which aligns with the experimental annotation from the DMS\_bin\_score. Interestingly, other predictive models, such as AlphaMissense, ESM1b, and GPN-MSA, classified the mutation as benign. Previously, ClinVar had classified this mutation as "likely pathogenic," but a recent update reclassified it as a variant with Conflicting Interpretations of Pathogenicity. In contrast, dbSNP still categorizes this mutation as "likely pathogenic."

To analyze the structural implications of this mutation, we retrieved the protein structure from AlphaFold’s website and visualized both the wild-type and mutated sequences using PyMOL[25].

As shown in Fig.6, the mutation may lead to the formation of H-bond between N atom of Gln563 and the neighboring alpha helix. In contrast, the wild-type Glu563 cannot form H-bond due to the negatively charged carboxylate of Glu563. This additional interaction between the two alpha helices in the Glu563Gln mutant may reduce the flexibility of the protein’s tertiary structure, potentially altering its function.

Further support for the pathogenicity of the E563Q mutation comes from a study by Johnston et al. [13] on LZTR1 variants and their role in Noonan syndrome. The study examines a family with the E563Q mutation, where both parents, heterozygous for the mutation, showed no significant phenotypes. However, their two homozygous children

displayed severe manifestations of Noonan syndrome. The first child was diagnosed with biventricular hypertrophic cardiomyopathy (HCM) at birth and had distinct facial features, mild short stature, and pectus excavatum. He developed acute lymphoblastic leukemia at age 3 and is now in remission. His younger brother had an atrioventricular septal defect (AVSD), severe biventricular HCM, and a sacral meningocele, and died on day 4 from an inoperable cardiac defect.

These findings, coupled with our structural analysis, suggest that the E563Q mutation in LZTR1 is likely pathogenic, as our model predicted, contrary to the existing state-of-the-art models.

### 3.4.2 Case study of KAT6A

For the second case study, we investigated the E221K mutation in the KAT6A protein. KAT6A is a histone acetyltransferase responsible for acetylating lysine residues in H3 and H4 histones. As part of the MOZ/MORF complex, KAT6A exhibits histone H3 acetyltransferase activity. It also serves as a transcriptional coactivator for RUNX1 and RUNX2, and acetylates p53/TP53, controlling its transcriptional activity via association with PML.

Our model classified the E221K mutation as benign, which aligns with the DMS\_bin\_score. Clinvar currently classifies the variant as VUS while the state-of-the-art computational models predict it as pathogenic except for Alphamissense which labels it as ambiguous.

The protein does not have a consensus structure, and the AlphaFold predictions for the protein show low overall confidence. However, the region surrounding the E221K mutation lies within a high-confidence zone and is located in a coil structure. PyMOL visualization of both WT and mutated residues shows no interaction with neighboring residues (Fig.7), suggesting minimal impact on the protein's overall structure.

The E221 residue lies within two functional domains. The first one is a Zinc Finger (ZF) (residues in 206-265). The PRU00146 ZF has no defined function. Zinc fingers typically involve cysteine or histidine residues, while this mutation involves glutamic acid to lysine, which suggests a low likelihood of impacting ZF function. The second domain is an Interaction region with PML (promyelocytic leukemia) containing residues 144-664 [23]. However, structural visualization reveals that the E221K mutation is buried within the protein core and not exposed, making it less likely to participate in significant interactions with adjacent residues [16].

ClinVar initially classified this variant as likely benign but later reclassified it as VUS. According to the authors' submission on ClinVarMiner, this variant has not been reported in individuals affected with KAT6A-related conditions, and the advanced modeling of protein structure and biophysical properties such as structural, functional, and spatial information, amino acid conservation, physicochemical variation, residue mobility, and thermodynamic stability indicate that this missense variant is not expected to disrupt KAT6A protein function. This statement reinforces the hypothesis that such mutation may have a lower impact on the protein function as loops are disordered and may contribute less to the protein function, especially since the mutation's position in 221 does not belong to a functional region.

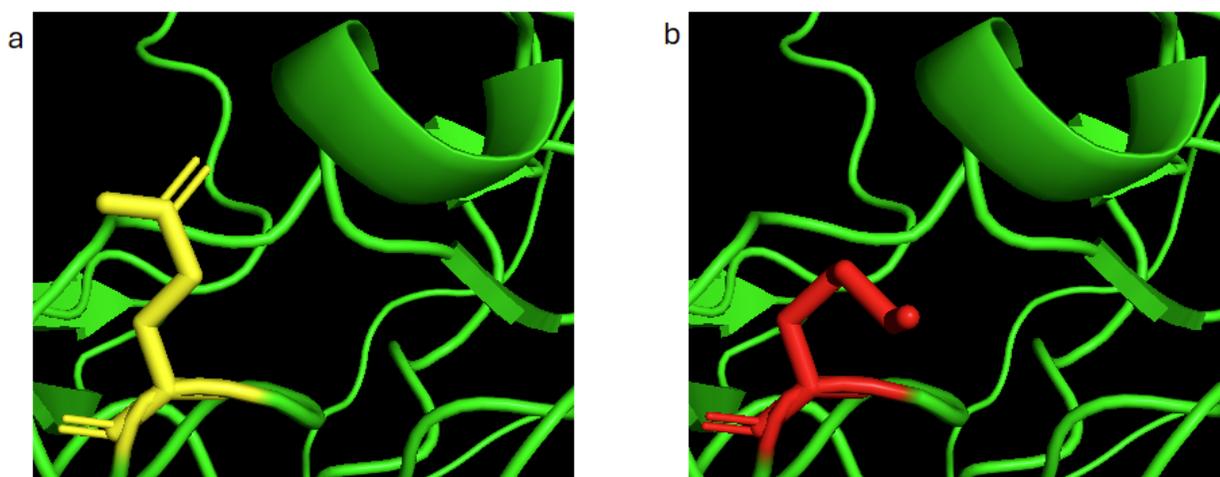


Figure 7: **Visualization of the E221K mutation in the KAT6A protein.** (a): The wild-type residue (E221) is depicted in yellow, showing no interactions with neighboring structures. (b): The mutated residue (K221) is shown in red, located in a coil region with no interaction with adjacent structures.

## 4 Conclusion

This study highlights the potential of combining advanced machine learning models in classifying genetic variants. Our integrated approach consistently outperformed the latest methods in determining the pathogenicity of these variants. The comprehensive evaluation clearly demonstrated the proficiency of the Multi-input Neural Network model in handling both straightforward cases and Variants of Uncertain Significance (VUS).

Our analysis underscored the importance of feature selection. Combining DNA and protein data using potential scores from GPN-MSA and ESM1b established a solid performance baseline. Adding AlphaMissense’s scores significantly boosted the predictive power of all models. This validates the advantage of integrating structural insights from protein data with sequence-based predictions. Using both observed and potential scores together led to the best overall results. The hypothesis that emphasizing observed scores would enhance the model’s focus on clinically validated mutations proved to be successful. By adding weight to these observed scores, the model better captured real-world genetic variations, leading to improved accuracy and robustness in classification.

Our integrated model consistently showed strong performance across various testing scenarios, demonstrating its capability to effectively interpret complex datasets. ESM1b and AlphaMissense consistently outperformed GPN-MSA. This can be due to the fact ESM1b and AlphaMissense are protein-based models whereas GPN-MSA is trained on DNA data, likely because protein data provides critical structural and functional context necessary for accurate variant classification. The improvements seen when excluding variants classified as ambiguous by AlphaMissense, and the corresponding drop in accuracy for those variants, further emphasize the challenges that uncertain classifications pose to predictive tasks.

In the comparison with ClinVar’s annotations, the extended evaluation highlighted the efficiency of our integrated model in distinguishing variant pathogenicity across both the ProteinGym and ClinVar datasets, which are meticulously curated for experimental validity. The model’s robust performance, particularly in handling VUS, makes it potentially useful for clinical and research applications where accurate interpretation of ambiguous genetic data is crucial.

Case studies further showcased the practical value of our model’s predictions. Investigating protein structures and reviewing related literature supported the accuracy of our model and the DMS\_bin\_score annotations, underlining the model’s real-world applicability. This also indicates that our model was effectively trained, capturing underlying information from state-of-the-art models rather than merely replicating their predictions.

our model is a step forward for characterizing variants of unknown significance and paves the way for identifying new therapeutic targets (or better characterization) or improving models that use NGS data [10, 22]. Looking ahead, it is essential to extend our validations by testing the integrated model on larger and more diverse datasets. Incorporating additional relevant components, such as transcriptomics scores from SpliceAI[12], could further enhance the model’s performance. By integrating these scores, we can add another dimension to our model, combining genomic, proteomic, structural, and transcriptomic data, leading to even more accurate and reliable predictions.

## 5 Acknowledgements

We would like to thank the members of the IT team, Jonathan Schmiedt and Thomas Wursten for their technical assistance during this project. We would also like to thank Maria Eugenia Riveiro for bringing her invaluable expertise and thorough insights.

## References

- [1] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems*. arXiv:1603.04467 [cs]. Mar. 2016. DOI: 10.48550/arXiv.1603.04467. URL: <http://arxiv.org/abs/1603.04467> (visited on 06/24/2024).
- [2] Ivan Adzhubei, Daniel M. Jordan, and Shamil R. Sunyaev. “Predicting functional effect of human missense mutations using PolyPhen-2”. eng. In: *Current Protocols in Human Genetics* Chapter 7 (Jan. 2013), Unit7.20. ISSN: 1934-8258. DOI: 10.1002/0471142905.hg0720s76.
- [3] Gonzalo Benegas et al. *GPN-MSA: an alignment-based DNA language model for genome-wide variant effect prediction*. Oct. 2023. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10592768/#S4> (visited on 04/24/2024).

- [4] Nadav Brandes et al. “Genome-wide prediction of disease variant effects with a deep protein language model”. en. In: *Nature Genetics* 55.9 (Sept. 2023). Publisher: Nature Publishing Group, pp. 1512–1522. ISSN: 1546-1718. DOI: 10.1038/s41588-023-01465-0. URL: <https://www.nature.com/articles/s41588-023-01465-0> (visited on 04/24/2024).
- [5] Leo Breiman. “Random Forests”. en. In: *Machine Learning* 45.1 (Oct. 2001), pp. 5–32. ISSN: 1573-0565. DOI: 10.1023/A:1010933404324. URL: <https://doi.org/10.1023/A:1010933404324> (visited on 07/11/2024).
- [6] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. arXiv:1603.02754 [cs]. Aug. 2016, pp. 785–794. DOI: 10.1145/2939672.2939785. URL: <http://arxiv.org/abs/1603.02754> (visited on 05/14/2024).
- [7] Jun Cheng et al. “Accurate proteome-wide missense variant effect prediction with AlphaMissense”. In: *Science* 381.6664 (Sept. 2023). Publisher: American Association for the Advancement of Science, eadg7492. DOI: 10.1126/science.adg7492. URL: <https://www.science.org/doi/10.1126/science.adg7492> (visited on 04/24/2024).
- [8] Kirsley Chennen et al. “MISTIC: A prediction tool to reveal disease-relevant deleterious missense variants”. eng. In: *PloS One* 15.7 (2020), e0236962. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0236962.
- [9] Shawn Fayer et al. “Closing the gap: Systematic integration of multiplexed functional data resolves variants of uncertain significance in BRCA1, TP53, and PTEN”. English. In: *The American Journal of Human Genetics* 108.12 (Dec. 2021). Publisher: Elsevier, pp. 2248–2258. ISSN: 0002-9297, 1537-6605. DOI: 10.1016/j.ajhg.2021.11.001. URL: [https://www.cell.com/ajhg/abstract/S0002-9297\(21\)00411-0](https://www.cell.com/ajhg/abstract/S0002-9297(21)00411-0) (visited on 06/05/2024).
- [10] Gabriele Fossi et al. *SwiftDossier: Tailored Automatic Dossier for Drug Discovery with LLMs and Agents*. 2024. arXiv: 2409.15817 [cs.AI]. URL: <https://arxiv.org/abs/2409.15817>.
- [11] Douglas M. Fowler and Stanley Fields. “Deep mutational scanning: a new style of protein science”. en. In: *Nature Methods* 11.8 (Aug. 2014). Publisher: Nature Publishing Group, pp. 801–807. ISSN: 1548-7105. DOI: 10.1038/nmeth.3027. URL: <https://www.nature.com/articles/nmeth.3027> (visited on 05/14/2024).
- [12] Kishore Jaganathan et al. “Predicting Splicing from Primary Sequence with Deep Learning”. English. In: *Cell* 176.3 (Jan. 2019). Publisher: Elsevier, 535–548.e24. ISSN: 0092-8674, 1097-4172. DOI: 10.1016/j.cell.2018.12.015. URL: [https://www.cell.com/cell/abstract/S0092-8674\(18\)31629-5](https://www.cell.com/cell/abstract/S0092-8674(18)31629-5) (visited on 05/15/2024).
- [13] Jennifer J. Johnston et al. “Autosomal Recessive Noonan Syndrome Associated with Biallelic LZTR1 Variants”. In: *Genetics in medicine : official journal of the American College of Medical Genetics* 20.10 (Oct. 2018), pp. 1175–1185. ISSN: 1098-3600. DOI: 10.1038/gim.2017.249. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6105555/> (visited on 06/05/2024).
- [14] Martin Kircher et al. “A general framework for estimating the relative pathogenicity of human genetic variants”. eng. In: *Nature Genetics* 46.3 (Mar. 2014), pp. 310–315. ISSN: 1546-1718. DOI: 10.1038/ng.2892.
- [15] Melissa J. Landrum et al. “ClinVar: public archive of interpretations of clinically relevant variants”. eng. In: *Nucleic Acids Research* 44.D1 (Jan. 2016), pp. D862–868. ISSN: 1362-4962. DOI: 10.1093/nar/gkv1222.
- [16] B. Lee and F.M. Richards. “The interpretation of protein structures: Estimation of static accessibility”. In: *Journal of Molecular Biology* 55.3 (1971), 379–IN4. ISSN: 0022-2836. DOI: [https://doi.org/10.1016/0022-2836\(71\)90324-X](https://doi.org/10.1016/0022-2836(71)90324-X). URL: <https://www.sciencedirect.com/science/article/pii/002228367190324X>.
- [17] Heng Li. “Tabix: fast retrieval of sequence features from generic TAB-delimited files”. In: *Bioinformatics* 27.5 (Mar. 2011), pp. 718–719. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq671. URL: <https://doi.org/10.1093/bioinformatics/btq671> (visited on 05/15/2024).
- [18] Zeming Lin et al. *Language models of protein sequences at the scale of evolution enable accurate structure prediction*. en. Pages: 2022.07.20.500902 Section: New Results. July 2022. DOI: 10.1101/2022.07.20.500902. URL: <https://www.biorxiv.org/content/10.1101/2022.07.20.500902v1> (visited on 06/25/2024).
- [19] Pauline C. Ng and Steven Henikoff. “SIFT: Predicting amino acid changes that affect protein function”. eng. In: *Nucleic Acids Research* 31.13 (July 2003), pp. 3812–3814. ISSN: 1362-4962. DOI: 10.1093/nar/gkg509.
- [20] Pascal Notin et al. *ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction*. Dec. 2023. URL: <https://www.biorxiv.org/content/10.1101/2023.12.07.570727v1.full> (visited on 04/24/2024).

- 
- [21] Dahui Qin. “Next-generation sequencing and its clinical application”. In: *Cancer Biology & Medicine* 16.1 (Feb. 2019), pp. 4–10. ISSN: 2095-3941. DOI: 10.20892/j.issn.2095-3941.2018.0055. URL: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6528456/> (visited on 06/24/2024).
- [22] Salvatore Raieli et al. *Escaping the Forest: Sparse Interpretable Neural Networks for Tabular Data*. 2024. arXiv: 2410.17758 [cs.LG]. URL: <https://arxiv.org/abs/2410.17758>.
- [23] Susumu Rokudai et al. “MOZ increases p53 acetylation and premature senescence through its complex formation with PML”. In: *Proceedings of the National Academy of Sciences* 110.10 (Mar. 2013). Publisher: Proceedings of the National Academy of Sciences, pp. 3895–3900. DOI: 10.1073/pnas.1300490110. URL: <https://www.pnas.org/doi/10.1073/pnas.1300490110> (visited on 06/21/2024).
- [24] Ashish Vaswani et al. *Attention Is All You Need*. arXiv:1706.03762 [cs]. Aug. 2023. DOI: 10.48550/arXiv.1706.03762. URL: <http://arxiv.org/abs/1706.03762> (visited on 05/14/2024).
- [25] Shuguang Yuan, H. Chan, and Zhenquan Hu. “Using PyMOL as a platform for computational drug design”. In: *Wiley Interdisciplinary Reviews: Computational Molecular Science* 7 (Jan. 2017), e1298. DOI: 10.1002/wcms.1298.