University of Glasgow | School of Computing Science

Extracting the Genotype of Mice Samples from Public Bioinformatics Database Using BioNER Techniques

Dhruv Kumar Patwari

School of Computing Science
Sir Alwyn Williams Building
University of Glasgow
G12 8QQ

A dissertation presented in part fulfilment of the requirements of the Degree of Master of Science at The University of Glasgow

1st September 2023

Abstract

The field of biomedical research has seen remarkable progress, leading to a significant increase in biological data stored in public databases like NCBI. These databases hold the potential for discovering new medicines and identifying potential risks to human health. However, the data within these repositories is often disorganized. Metadata in the repositories is incomplete and inconsistent. Much of the valuable information is stored in plain text, making it difficult for researchers to gain new insights. One important type of information is genotypic data, which can provide valuable insights into the inheritance of traits, disease susceptibility, and intricate molecular interactions within living organisms. This dissertation focuses on identifying genotypes mentioned in project descriptions. This is achieved by creating a collection of descriptions that already contain known genotypic information. These description-genotype pairs are then used to train various BERT models. The process of identifying genotypes proved to be a challenging task. Among the different methods we tried, curriculum learning using BioRED was the most successful. While the model's accuracy is imperfect, it provides researchers with structured data that can be further verified manually. In this dissertation, we aim to improve the extraction of genotypic information from the complex and unorganized realm of bioinformatics databases.

Education Use Consent

I hereby give my permission for this project to be shown to other University of Glasgow students and to be distributed in an electronic format. **Please note that you are under no obligation to sign this declaration, but doing so would help future students.**

Name: Dhruv Kumar Patwari Signature: 

Acknowledgements

I extend my heartfelt gratitude to my supervisor, Dr. Jake Lever, for his invaluable guidance and mentorship that have been instrumental in steering this project. Exploring the vast realm of the biomedical domain has been an enriching experience. I also sincerely appreciate my family's unwavering support during this dissertation. Additionally, I would like to thank Hrithika for her continuous support and contributions.

Contents

Chapter 1: Introduction

In biomedical research, data holds a significance akin to that of oil. Just as oil fuels industries, high-quality data fuels rapid discoveries in medicine. This accelerated discovery process can rescue countless lives and elevate societies' overall health and wellness. To expedite this transformative journey, biomedical researchers embrace openly sharing their research data via expansive platforms like the National Center for Biotechnology Information's (NCBI) BioSample, BioProject [7], and Sequence Read Archive (SRA) [26]. These platforms house extensive samples and sequences drawn from millions of experiments, thus imbuing them with immense value. However, despite the promising outlook of these databases, the research community encounters challenges in fully harnessing their boundless potential. The intricate process of extracting meaningful insights from these databases presents hurdles that must be overcome.

1.1 Motivation

Because of their vast coverage across various biological domains, navigating and retrieving relevant information from public repositories like NCBI has proven to be challenging [20]. The absence of centralized catalogues exacerbates this issue, raising the barriers to entry. Furthermore, the exponential growth of data strains the already outdated and overburdened infrastructure [9].

Compounding these challenges is the lack of consistency in data formats stored within these databases. This inconsistency hinders the integration of data, which could otherwise offer valuable insights. Compromised metadata and inconsistent naming conventions further complicate matters [18]. Crafting queries specific to each database necessitates domain expertise to glean meaningful insights from the data. [8]

Given the obstacles above, cleaning and structuring data uniformly becomes imperative, ensuring its accessibility for future experiments. However, achieving this task manually is unfeasible, thus prompting the adoption of Natural Language Processing (NLP) techniques. These techniques have proven beneficial in addressing the issues with biomedical texts.

Within biomedical research, a critical application of NLP is Biomedical Named Entity Recognition (BioNER). BioNER automates the identification of specific entities in medical texts, eliminating the need for manual intervention. A significant entity in this context is "Genotype," a term of paramount importance as it provides insights into genetics and diseases.

Genotypes unveil the inheritance of traits, susceptibility to diseases, and intricate molecular interactions within living organisms. Despite their significance, there remains substantial unexplored territory concerning genotypes within the scope of BioNER.

This research project aims to create a procedure for extracting Genotypes from biomedical texts. This procedure can also be adapted to extract other lesser-explored metadata attributes. This enhancement in metadata quality is expected to make a valuable contribution to the advancement of biomedical research.

1.2 Aim and Objectives

BioNER, unlike standard NER, poses intricacies due to ambiguous entity boundaries and varied representations. Researchers have proposed methods (outlined in the subsequent chapter), but uncharted territories remain.

This project seeks to cultivate a model adept at discerning genotypes within medical texts. The model will be trained using a custom dataset with automated annotations. Performance evaluation will hinge on span-level F1 score.

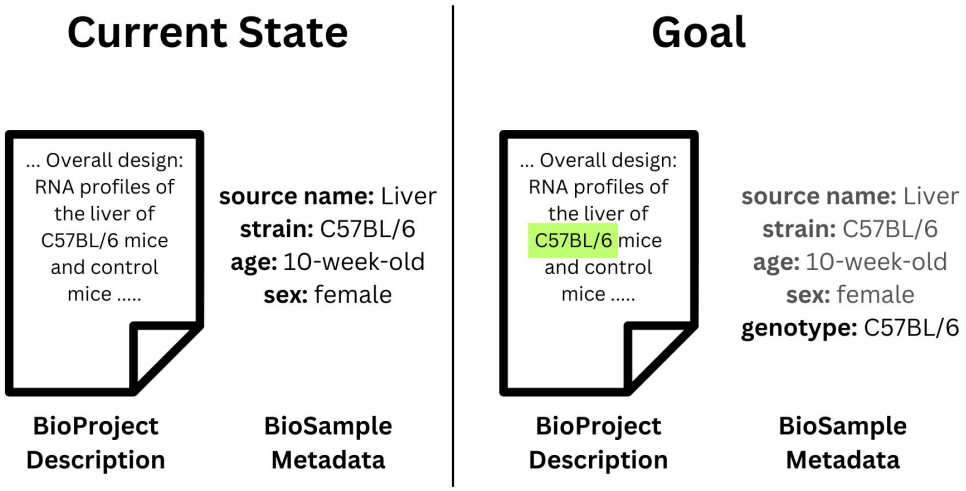

Figure 1.1: This project aims to find the genotypes in the project description so the researchers can use them to make new observations. Text from [2]

The project's objectives are as follows:

1. **Survey Cutting-Edge BioNER Techniques:** Conduct a comprehensive literature review to comprehend the current state-of-the-art BioNER techniques, elucidating their strengths and weaknesses.
2. **Construct an Annotated Database:** Develop a repository tailored for model training using the already available BioSamples with genotypes explicitly mentioned.
3. **Engineer and Refine BERT Models:** Engineer BERT models and meticulously fine-tune them on the curated dataset, enhancing their proficiency in genotype recognition.
4. **Methodically Evaluate Model Performance:** Rigorously assess model performance, comparing results to ascertain the most effective approach.

1.3 Report Structure

The report contains six chapters. Beginning with an introduction that explains the project's motivation, the survey chapter delves into the necessary background and related literature, establishing the project's context. The chapter on dataset creation outlines how data was collected, structured, and labelled. The subsequent chapter provides detailed insights into the model architecture, including setup, tuning, and integration of BioRED data. The evaluation

chapter rigorously addresses research questions by assessing model performance using span-level F1 scores. The conclusion summarises achievements, suggests future directions, and encapsulates the project's contributions to biomedical NLP.

Chapter 2: Survey

2.1 Background

2.1.1 Biomedical repositories

The NCBI repository is a prominent biomedical data hub, consolidating diverse biological data like sequences, structures, and pathways. This centralised repository offers extensive data access and is pivotal in modern biological research.

Bioinformatics databases can be classified into primary and secondary categories. Primary Databases encompass raw experimental data, including sequences and structures, while Secondary Databases house analysed and curated information derived from primary sources and literature.

However, challenges hinder their broader utility. These include data interoperability, proprietary limitations, standardisation issues, complex interfaces requiring expertise, user proficiency demands, and data quality concerns [8]. Overcoming these obstacles necessitates collaboration among biologists, computer scientists, curators, and developers to optimise the potential of bioinformatics databases.

2.1.2 BioProject and BioSample Databases

In 2011, NCBI introduced BioProject and BioSample to streamline the arrangement and categorisation of project metadata and sample details submitted to NCBI and other repositories. BioProject is a centralised platform for delineating research projects and establishing connections with associated data in NCBI's archival databases, including SRA, dbGaP, and GenBank. On the other hand, BioSample captures comprehensive information about biological specimens utilised in experiments, establishing connections with corresponding projects and experimental data. Collectively, BioProject and BioSample enhance the capacity to query, locate, integrate, and interpret the extensive volume of data within NCBI's archives [7].

Each BioProject has a project description, metadata and one or more BioSamples. These BioSamples have more useful metadata tagged by the researchers. As shown in Figure 2.1, the BioProject has a long description and various attributes. This project has multiple samples; an example of one of the samples can be seen in Figure 2.2.

2.1.3 BioRED

BioRED serves as a substantial biomedical relation extraction dataset, encompassing annotations for various entity categories such as gene/protein, chemical, variant, disease, species, and cell line, as well as relation pairs (e.g. gene-disease, chemical-chemical) across a collection of 600 PubMed abstracts.

This dataset was conceived to stimulate the advancement of general biomedical relation extraction systems, extending beyond single-entity or relation-type constraints. Its provision of multi-entity and multi-relation annotations in a unified dataset catalyses this objective.

Figure 2.1: BioProject data on NCBI [3]

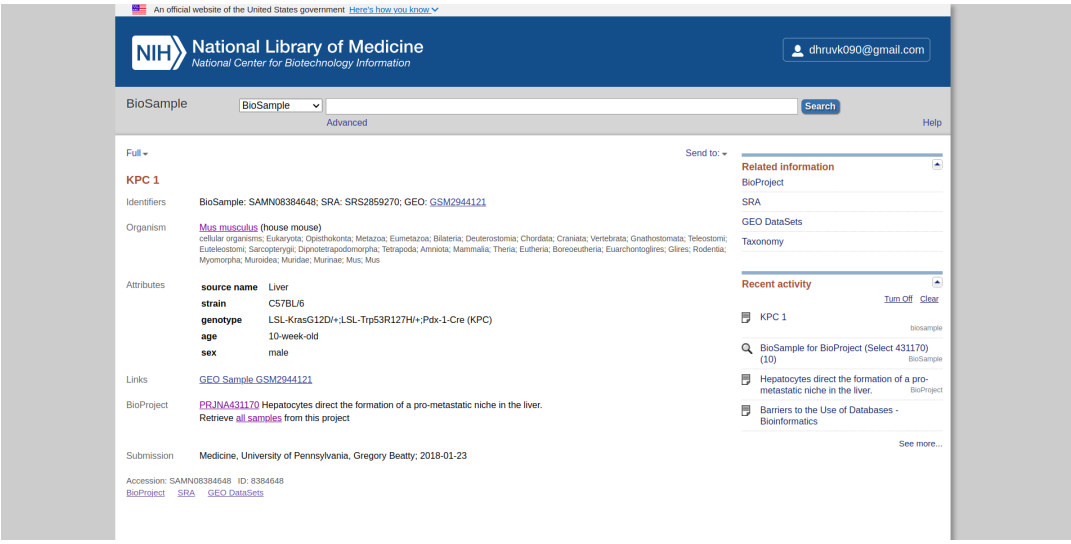

Figure 2.2: One of the samples in the BioSample database from the project. [5]

An intriguing challenge lies in document-level relation annotation, which presents more complexity than sentence-level annotation. Additionally, the dataset introduces a novelty detection capability, enabling the differentiation between novel and background knowledge.

BioRED is rich in information, featuring 20,419 entity mentions encompassing 3,869 unique identifiers and 6,503 annotated relations. Notably, 69% of these relations are identified as novel findings. On average, each abstract comprises 11.9 sentences and 304 tokens. An average abstract is annotated within this context with 10.8 relations and 34 entity mentions (3.8 unique).

2.2 Literature Survey

2.2.1 Metadata

Metadata serves as indispensable contextual information, enabling the meaningful comprehension and utilization of data housed within public biomedical databases [17]. Some

metadata attributes include genotypes, age, experimental protocols, and database cross-references. Experimental protocols, encompassing intricate insights into library preparation, instrumentation, and analysis methods, are paramount in guaranteeing research reproducibility [22]. Additionally, database cross-references are pivotal in establishing linkages among correlated records across diverse resources, facilitating effective data integration [7].

The growing amount of data in these databases underscores the importance of extensive metadata coverage to improve the effectiveness of data queries. However, a significant challenge these databases face is insufficient metadata coverage. Specifically, in the case of SRA BioSample, essential attributes like age, genotype, and protocol often lack annotations, significantly limiting the potential for reusing data and conducting comprehensive meta-analyses [22].

Multiple endeavours have been undertaken to augment metadata coverage and accessibility. This encompasses the streamlining of the submission process through templates [37], the application of deep learning-based named entity recognition to extract metadata from free-text [22], the automatic suggestion of metadata using established ontologies [35], and the utilization of metadata from publications to populate database entries [15].

Beyond incomplete metadata, heterogeneity represents a substantial challenge. User-defined fields and non-standardized terms hamper interoperability [37]. Initiatives such as MIxS introduce standardized metadata packages tailored to genomes, metagenomes, and other data types [6]. Efforts involving controlled vocabularies, ontologies, and semantic annotations strive to enhance consistency and enable seamless integration across databases [37]. Although curating and harmonizing metadata necessitate substantial manual effort, community engagement in annotation and standardization is pivotal for ongoing enhancements [30].

Concurrent with metadata generation, the availability of practical retrieval tools is paramount for researchers. Resources like ffq enable efficient querying of metadata linked to NCBI accessions and facilitate direct access to raw data [15]. The availability of robust metadata fosters interpretation and encourages the reuse of archived data.

2.2.2 BioNER

Named entity recognition (NER) involves identifying specific named entities in text and categorizing them into predefined groups, such as individuals, locations, organizations, etc. When applied to the biomedical field, this task is called Biomedical Named Entity Recognition (BioNER). BioNER aims to detect specialized biomedical entities such as genes, proteins, diseases, chemicals, drugs, species, mutations, and so forth in unstructured biomedical texts. It plays a pivotal role in extracting valuable information from biomedical texts and conducting text analysis. [10]

BioNER is regarded as more intricate than general NER due to challenges unique to medical texts. These texts lack consistent naming conventions [29], new entities can frequently emerge (e.g., COVID-19) with a mix of numbers, letters, and symbols, necessitate substantial domain expertise [29], and often lack extensive labelled data [21].

Traditionally, BioNER relied on rule-based [23], dictionary-based (utilizing ontology lookup tables) [28], and feature-based systems. The latter encompassed machine learning (ML) techniques such as Support Vector Machines (SVM) [25] or Conditional Random Fields (CRF) [33], particularly as the volume of data increased. However, these methods demanded specialized biomedical domain knowledge [29], limiting accessibility, and their performance was restricted in the face of novel entities in swiftly expanding medical texts.

Recent Deep Learning (DL) advances have provided more accessible solutions for addressing this challenge. Contemporary BioNER systems are founded on neural networks, particularly Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer networks. CNN-based models employ convolutional filters to extract character-level features to handle various word forms [11]. RNNs, like Long Short-Term Memory (LSTM) networks, capture long-distance relationships in biomedical text, and bidirectional RNNs encompass both past and future context [16]. Transformer-based models, exemplified by BERT [12], capture contextual associations irrespective of distance. Integration with CRF for structured prediction enhances these models' BioNER performance by autonomously learning features from data [19].

Innovations in the BioNER domain encompass Pretrained Language Models like BioBERT [24], BlueBERT [31], and PubMedBERT [14] tailored to biomedical text. Additionally, innovations include the incorporation of syntactic features like Part-of-Speech (POS) tags and parse trees through techniques like key-value memory networks (KVMN) [36], multi-task learning across diverse entity types and datasets [4], data augmentation strategies [13], noise reduction methods for annotations [38], techniques to mitigate model biases and enhance generalization [4], and knowledge-based approaches that leverage ontologies [27].

2.3 Problem Statement

Through an analysis of research within the BioNER domain, it becomes evident that there is potential to expand the identification of attributes within metadata. However, several challenges impede progress in this endeavour. Primarily, attributes less extensively studied, such as genotypes, lack dedicated databases. Moreover, the effective training of models for these attributes necessitates access to pertinent databases to yield satisfactory performance viable for researchers. In addressing these challenges, the present dissertation project strives to establish a new genotype dataset, train existing deep-learning models, and evaluate their efficacy.

Another aspect we're exploring in this project is the potential benefit of training a model on a related but more straightforward task before tackling the main task. In this instance, we plan to train the BERT model on the gene dataset from BioRED before moving on to our dataset. This approach aligns with the concept of curriculum learning [34], where the model gradually builds its skills by starting with more accessible examples before advancing to more complex ones.

The main aim is to provide researchers with a helpful way to train models for identifying different entities in biomedical texts, especially when databases don't exist. By creating a complete dataset and training Deep Learning models, this project offers a valuable tool to help researchers effectively extract essential information from biomedical texts.

2.4 Research Objective

1. Objective 1: Create a dataset by matching genotypes to the project description.
2. Objective 2: Identify which tagging method works the best.
    - The entities need to be recognised in the text and then labelled using the BIO method (explained in Chapter 3)
    - Two tagging methods were identified: token matching and position matching labelling methods. (Explained in chapter 3)

3. Objective 3: Identify the model which performs the best on the dataset.
    - We train BERT, BioBERT and PubMedBERT to compare the performances.
4. Objective 4: Does curriculum learning help improve the best-performing model's performance?
    - We used Gene data for NER from the BioRED dataset to fine-tune the best-performing model to compare the performances.

Chapter 3: Dataset Creations

Currently, there is a lack of available datasets for recognising genotypes as named entities. Hence, one of the main objectives of this project was to establish a dataset tailored for this purpose. This was accomplished by harnessing the structured data in the biosamples database, with a specific focus on genotypes related to mice. These BioSamples were paired with their corresponding BioProjects. The resulting database was subsequently utilised to train a variety of models.

3.1 Tools Used

The project was completed using Python, PyTorch, sklearn, and Bio Package to access Entrez.

The hardware used for training was Google Colab's free version, Kaggle Notebooks and the IDA GPU cluster provided by the University.

3.2 Downloading Data

3.2.1 Libraries and Packages used

BioProject and BioSample databases, due to their significant size, pose challenges in downloading and filtering data comprehensively. To mitigate this, NCBI offers an effective querying method, the Entrez Programming Utilities (E-utilities). These utilities comprise nine server-side programs, furnishing a stable interface to the NCBI's Entrez query system. Using a consistent URL syntax, the E-utilities transform input parameters to facilitate data retrieval from 38 biomedical databases. This includes nucleotide sequences, gene records, and biomedical literature. By allowing software to send URLs to the E-utilities server, interpret XML responses, and create customised data pipelines, these utilities empower efficient data manipulation using diverse programming languages like Perl, Python, Java, and C++. [32]

For Python-based access to the Entrez query system, the Biopackage library is pivotal. It is utilising `Bio.Entrez`, this library permits querying various databases within the Entrez system, such as BioProject. It offers functions like `efetch`, enabling the retrieval of complete records from Entrez databases, and `einfo`, which provides an overview of accessible Entrez databases. This comprehensive toolset facilitates search operations, retrieval, and filtered data exploration. [1]

3.2.2 Retrieving the data

Employing the Biopython package facilitates targeted data retrieval, focusing specifically on projects involving mice in their research. The query is detailed in Figure 3.1. This query led to the identification of 49,527 projects aligned with the specified criteria. However, Entrez limits the retrieval to 10,000 records. Hence, we retrieve the first 10,000 projects from the filtered list. All the projects without BioSamples were immediately dropped before being saved to reduce compute time.

```
search_term = "Mus musculus[Organism]"
filters = "Primary submission[Filter] AND Transcriptome or Gene expression[Filter]"

handle = Entrez.esearch(
    db="bioproject", term=search_term + " AND " + filters
)
```

Figure 3.1: Query used to filter BioProject

Subsequently, using the project IDs from these projects, we queried and accumulated the samples associated with each project, compiling them into a separate data frame. 253,857 samples with 94 different metadata attributes matched the criteria. Among these samples, 115,118 have genotypes explicitly mentioned in their metadata, so we only used those samples.

The remaining samples were grouped according to their respective projects, and the genotypes referenced within each project were aggregated into a list format, serving as a distinct feature.

The grouped data frame was merged with the original BioProject data frame, employing project IDs as the linking attribute to enhance data organisation. Consequently, projects without genotypic information were removed from the dataset. A comprehensive data frame was constructed through these sequential operations, encompassing projects with explicit genotypic information within their metadata.

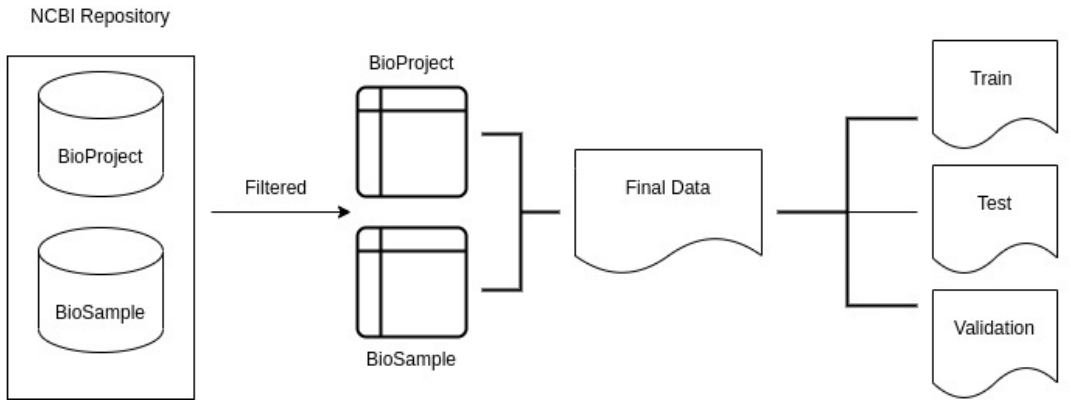

Figure 3.2: Framework of how the dataset was created

3.3 Structuring Data

Once we have the grouped dataset, we find the mentions of the genotypes in project descriptions. This process aims to identify and extract all instances where genotypes mentioned within the biosamples' metadata are in the project descriptions. Once we find a match, we store the location of the genotype as a new feature in the dictionary format where the key is one genotype, and the value is the list of tuples of start and end values of that genotype. Figure 3.5 is an example of the same.

Here, we ignore all the *wild-type* genotypes, *Knock out*, and *control* because they do not provide any additional information but skew the dataset, as they make up for over 30% of all the genotypes. This could hamper training the models.

We then drop all the projects that do not mention genotypes in their description. The resulting

```
ignore_genotypes = ['WT', 'wt', 'Wt', 'wild type', 'control', 'KO',
                    'wild-type', 'wildtype', 'Control', 'Wild type',
                    'Wild-type', 'Wildtype', 'Wild Type', 'knockout']
```

Figure 3.3: The genotypes that were ignored

data frame is the final dataset that we will use in the rest of our project.

3.4 Data Statistics

The genotypes were unevenly distributed, with 30% of them just being accounted for by the terms in Figure 3.3. To avoid any bias in the model, we decided to remove all the commonly occurring genotypes, like all variants of *wild type*, *knock out* and *control*.

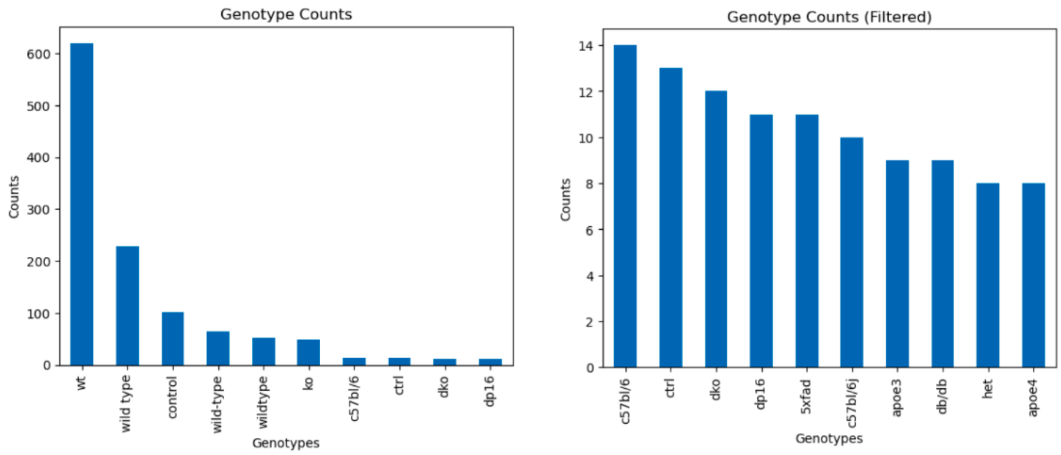

Figure 3.4: Top 10 genotypes by count. On the left is the initial distribution of the genotypes, and on the right is the distribution after removing the genotypes in the `ignore_genotypes` list

The project descriptions exhibit a range of sizes, from the shortest at 8 words to the longest at 742 words. On average, these descriptions consist of approximately 143 words.

The final data frame has 1804 records that exactly match the criteria. We then split the data into training, Validation and test sets, using the 60/20/20 split, which resulted in 1082 in training, 361 in validation and 361 in test set.

These were stored in the file system to preserve the distribution and ensure all the models were trained and tested on the same data.

3.5 Labelling

The Beginning-Inside-Outside (BIO) labelling scheme was employed in annotating project descriptions. This tagging approach is widely utilized in BioNER. It operates by designating the position of a token within a named entity, indicating whether the token is at the Beginning, Inside, or Outside of the entity.

The BIO labelling technique offers notable benefits. It provides a straightforward and intuitive method for annotating sequence labelling tasks, such as NER, streamlining the process

	ID	Title	Description	genotype	index_dict
0	986528	PIKfyve controls dendritic cell function and t...	To understand the mechanism of Pikfyve loss in...	[WT, KO]	{'WT': [(151, 153), (612, 614)], 'KO': [(193, ...
1	986137	Hdac1 and Hdac2 regulate the quiescent state a...	While cell division is essential for self-rene...	[Control, Hdac1/Hdac2 double mutant]	{'Control': [(1481, 1488)], 'Hdac1/Hdac2 doubl...
2	986078	Evi1 governs Kdm6b-mediated histone demethylat...	Ecotropic viral integration site 1 (EVI1/MECOM...	[WT, Evi1 overexpressing]	{'WT': [(603, 605), (1001, 1003)], 'Evi1 overe...
3	985683	Reduction of Nemo-like kinase increases lysoso...	Protein aggregation is a hallmark of many neur...	[wild-type, Nlk KO]	{'wild-type': [(1327, 1336)], 'Nlk KO': []}
4	985681	Mitochondrial PGAM5–Drp1 signaling regulates t...	Macrophages play a critical role in the regula...	[Control, PGAM5 KD]	{'Control': [(1999, 2006)], 'PGAM5 KD': [(2010...

Figure 3.5: Example of the dataset before tokenization.

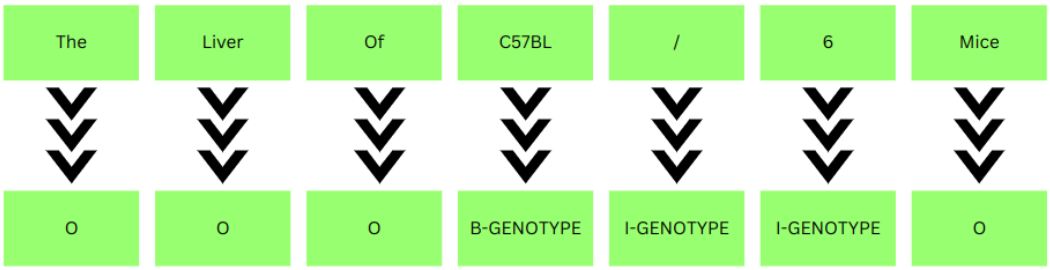

Figure 3.6: Example of how BIO Labels are assigned

and enhancing clarity in data annotation. Furthermore, the explicit demarcation of entity boundaries, a characteristic feature of the BIO labelling approach, proves particularly valuable in tasks like BioNER, contributing to the precise identification of entities within the text.

3.5.1 Token Matching

Recognizing genotypes within free text poses challenges for language models like BERT due to their potential to span across multiple tokens. This complexity makes accurate isolation from surrounding text challenging for these models. For instance, the genotype *LSL-KrasG12D/+;LSL-Trp53R127H/+;Pdx-1-Cre (KPC)* is extensive and may be segmented into multiple tokens by the model.

We implemented a strategy to enhance the learning process involving tokenising genotypes and descriptions. We applied BIO labels to ensure comprehensive annotation by aligning tokens between these elements. This methodology enabled us to accurately identify genotype mentions, even when presented in abbreviated forms, providing appropriate token-label correspondence. However, this approach did present challenges, as certain tokens unrelated to genotypes were incorrectly labelled as such, leading to an increase in the error rate. Figure 3.7 shows how token matching works on a project description.

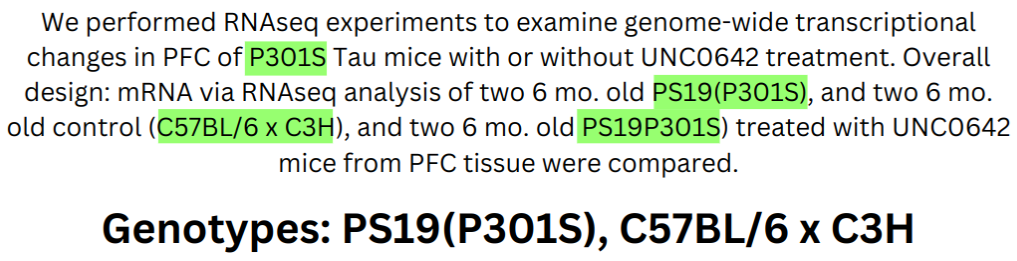

Figure 3.7: Example of how token matching labels genotypes in a project description [2].

3.5.2 Position Matching

We retained the specific positions of genotypes within the text by employing exact string matching. In this procedure, we tokenized the description and utilized the previously established offset mapping to accurately assign BIO labels to the complete and precise mentions of genotypes. This approach enabled us to effectively capture and label the entire genotype mentioned within the text. Figure 3.8 depicts how position matching works on a project description. It can be seen that only the exact match is recognized as a genotype. The partial genotype mentions are ignored by this method.

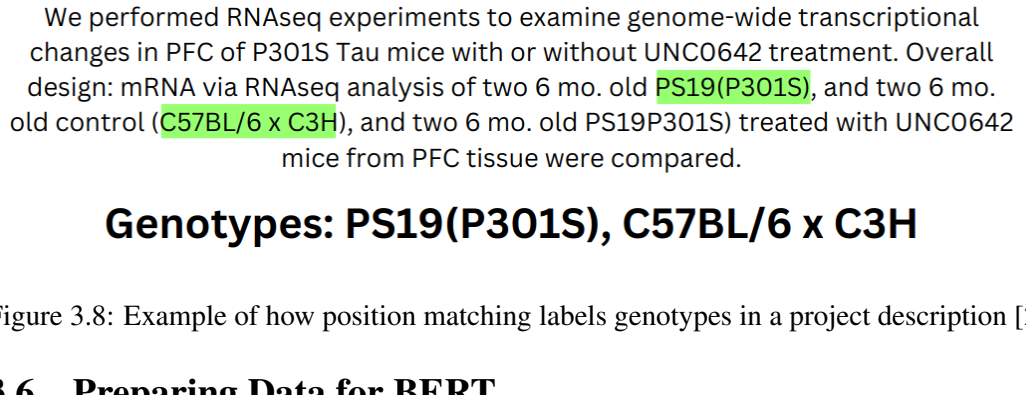

Figure 3.8: Example of how position matching labels genotypes in a project description [2].

3.6 Preparing Data for BERT

The upcoming chapter will delve into the discussion of various BERT models trained using the acquired data. However, the present dataset necessitates preprocessing and modifications before it becomes suitable for training these models. The data underwent three primary stages, which are outlined below.

Firstly, we employed a tokeniser, such as the `BertTokenizer` from the Hugging Face library, for text preprocessing and tokenization. This process involved breaking down the text into subwords or tokens. We restricted the input text to adhere to BERT's maximum input length of 512 tokens. Since descriptions often surpassed this limit, we divided each into up to 500 token segments. Tokens exceeding this limit were treated as separate rows in the data frame. This strategy was implemented to ensure the retention of all available information. These adjustments were implemented after the assignment of BIO labels using their respective techniques, streamlining the overall process.

Additionally, we converted tokens to input IDs and created attention masks. The tokenized text was transformed into input IDs, numerical representations of tokens based on the BERT vocabulary. An attention mask was crafted to indicate which tokens corresponded to actual words and which were padding tokens.

Subsequently, all the generated lists were used as inputs for the model to initiate the training and evaluation process.

Chapter 4: Models and Method

4.1 Model Overview

With the dataset in place, the subsequent step involves training various BERT [12] models to determine the most effective performer for our specific case. We will focus on two BERT model variations: BioBERT [24] and PubMedBERT [14]. These models' performances will be evaluated against the baseline vanilla BERT model to ascertain their respective strengths in our context.

4.1.1 BioBERT

The collaborative effort of researchers from NAVER, KAIST, and UCSD resulted in the development of BioBERT. Constructed through the progressive pretraining of the original BERT model using biomedical text, the pretraining corpus comprised 4.5 million PubMed abstracts and 13.5 million full-text articles from PubMed Central. The vocabulary employed adopts WordPiece tokenization, encompassing domain-specific terms inherited from the foundational BERT model. This adaptation demonstrates robust capabilities in various biomedical tasks, like NER, Question Answering (QA) and relation extraction [24].

4.1.2 PubMedBERT

Engineered by a team of researchers from Microsoft Research, PubMedBERT distinguishes itself by being trained exclusively on biomedical text. The pretraining process involved a substantial corpus of 14 million PubMed abstracts, amounting to 3 billion words. The vocabulary employed was meticulously constructed from biomedical text alone, using the WordPiece tokenization method. This adaptation showcases remarkable prowess, surpassing even BioBERT and other models when evaluated against the BLURB benchmark. Its exceptional performance extends across various biomedical Natural Language Processing (NLP) tasks [14].

4.2 Setup

The necessity for fine-tuning the BERT model for our specific task arises from its pre-trained nature on vast language texts, which does not inherently possess knowledge of solving our current problem. In this context, the task involves named entity recognition (NER), which demands incorporating a linear layer into the pre-trained BERT model to facilitate mapping to NER tags.

In this regard, PyTorch provides a convenient solution with the `BERTForTokenClassification` model. This model constitutes a BERT architecture and a singular linear layer tailored for classification tasks. Both the pre-trained model and our untrained layer collaboratively learn from the data fed into the system.

Before utilizing the data, tokenization is required, and for this purpose, the `BertTokenizer` is employed. Notably, the `BertTokenizer` offers various models to select from. Throughout this project, consistency was maintained by using the same tokenizer and model.

The data is presented in a composite form encompassing input IDs (tokens represented as IDs), attention masks, and labels. This amalgam of information serves as the input to the model, enabling it to undertake the process of learning and classification.

We selected the Adam optimizer for its advantages in fine-tuning our BERT models for NER tasks. The Adam optimizer offers faster convergence during training by combining momentum and RMSProp optimizers. It dynamically adjusts learning rates for each parameter based on gradient moments, which proves effective in managing varying learning rates for complex tasks like NER. Additionally, the inclusion of bias correction in Adam introduces a warm-up mechanism that reduces the learning rate at the beginning of training, contributing to the stability of the fine-tuning process.

In this project, we employed the `BERTForTokenClassification` class with various models: BERT (bert-base-uncased), BioBERT (dmis-lab/biobert-base-cased-v1.2), and PubMedBERT (microsoft/BiomedNLP-PubMedBERT-base-uncased-abstract). The setup remained consistent across these models, with the sole alteration being the pre-trained model chosen for evaluation. The BERT models used in this experimentation had a maximum token length of 512, and due to GPU limitations, the batch size was constrained to 8. The labels [PAD], O, I-GENOTYPE, and B-GENOTYPE correspondingly mapped to 0, 1, 2, and 3. For each model, the optimal batch size and learning rate were determined through grid search, and the configuration yielding the best performance was selected for extended training over multiple epochs.

4.3 Hyperparameter tuning

The careful selection of hyperparameters, such as learning rate, batch size, and epochs, profoundly influences the efficacy of BERT models. This underlines the significance of hyperparameter tuning, as it plays a pivotal role in attaining optimal performance from BERT models. An optimized configuration achieved through tuning is better suited to generalize over new, unseen data compared to default hyperparameters. Additionally, hyperparameter tuning mitigates the risk of overfitting by determining an appropriate model capacity for the given dataset. In the Named Entity Recognition (NER) context, improper hyperparameters can result in diminished F1 scores due to inadequate entity recognition.

To address this, we employed the `GridSearchCV` function from the Scikit-learn (SK-learn) library's `model_selection` package. This utility method systematically assesses all combinations of hyperparameters specified in a defined grid. This method identifies the optimal configuration within the search space by conducting an exhaustive search, enhancing the model's performance.

Due to resource constraints, our grid search was constrained to three discrete values for both the learning rate and batch size. Specifically, we examined the values 3e-5, 1e-5, and 5e-5 for the learning rate. We explored the 2, 4, and 8 alternatives regarding the batch size. This strategic limitation enabled us to efficiently navigate the hyperparameter space while optimizing the model's performance.

4.4 BioRED + BERT

We employed the BioRED dataset to preliminary fine-tun the best-performing model before adapting it to our specific task. By doing this, we provided contextual knowledge to the model before training it on our target problem. Our approach involves training the model on BioRED's gene data, serving as a preparatory step before training it on the primary task

of recognizing genotypes. This process follows a curriculum learning strategy, aiding the model's learning progression.

In genetics, a *gene* is a segment of DNA that codes for a specific function. In contrast, a *genotype* encompasses an individual's overall genetic composition, including the combination of alleles in their DNA.

We specifically extracted gene-related information from the comprehensive BioRED dataset for our analysis. Among the three models, we selected the top-performing one based on its F1 Score at the span level. We conducted NER on a subset of 400 gene mentions records with this chosen model. This involved the entire training and fine-tuning process, resulting in a model fine-tuned on the BioRED dataset. Subsequently, we employed this fine-tuned model to train on our current dataset to assess whether this approach influences the model's performance.

Data preprocessing, hyperparameter tuning, and fine-tuning for the model intended to be trained on the BioRED dataset mirrored the procedure above. Notably, this model underwent a dual training process on distinct datasets.

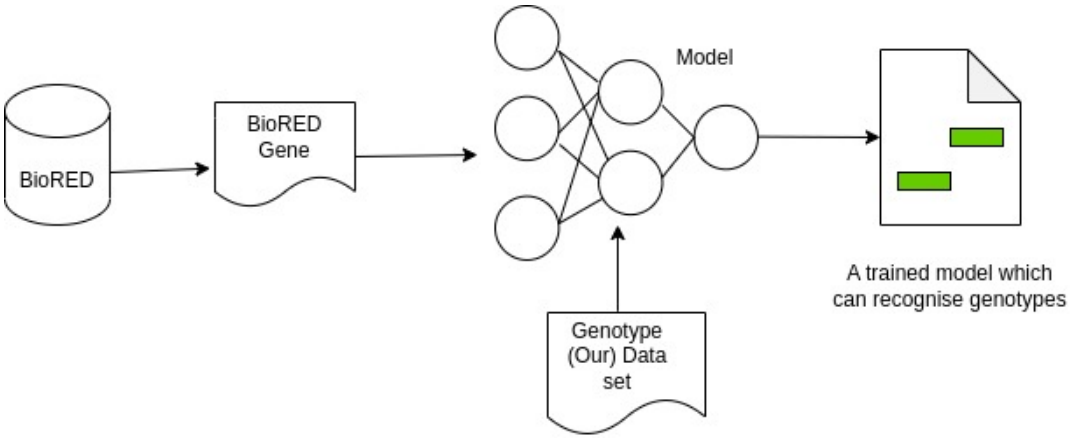

Figure 4.1: Architecture diagram showing how the model was trained.

4.5 Evaluation Criteria

4.5.1 Precision

Precision quantifies the proportion of accurate positive predictions made by a model. This metric assesses the accuracy of the model's positive forecasts, highlighting its ability to predict positive instances accurately. Precision is particularly valuable when minimizing the occurrence of false positives holds significance. It indicates the model's positive predictive value and higher values indicate more accurate positive predictions.

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4.1}$$

4.5.2 Recall

Recall gauges the proportion of positive instances the model correctly identifies as positive. This metric evaluates the comprehensiveness of the model's positive predictions, indicating how effectively it captures positive instances. Recall becomes valuable when the priority is

to minimize the occurrence of false negatives. It measures the model's sensitivity in capturing positive samples, with higher values indicating better performance in identifying actual positives.

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{4.2}$$

4.5.3 Spam level F1 score

The F1 score is an essential metric used to evaluate the performance of classification models. It provides a balanced measure by combining precision and recall through their harmonic mean.

The calculation involves generating tuples containing start and end indices of entities from both the true and predicted lists of labels. These tuples are then compared to compute the model's performance.

The F1 score evaluates a model's ability to identify named entities in a given text in NER tasks. This evaluation is conducted on a per-entity basis rather than focusing on individual tokens.

$$F1_{span} = \frac{2 * precision_{span} * recall_{span}}{precision_{span} + recall_{span}} \tag{4.3}$$

Chapter 5: Evaluation

5.1 Overview

This chapter focuses on the evaluation and analysis of the model's performance, with evaluations conducted to address the following research questions:

Research Question 1 (RQ1): Does the labelling and tagging method impact the model's performance?

Research Question 2 (RQ2): How does the performance of specialized models like BioBERT compare to that of a general model like BERT?

Research Question 3 (RQ3): To what extent does curriculum learning improve the model's performance?

5.2 Experimental Setup

All models underwent training and validation using identical training and validation sets to ensure consistency. The conclusive metrics guiding decision-making were generated from the test set, which remained completely unseen by the models.

5.3 RQ1: Does the labelling and tagging method impact the model's performance?

To address our initial research question, we compared two tagging methods to determine their respective performance levels. Maintaining model consistency with BERT, we initiated hyperparameter tuning for each configuration, employing 5 epochs per set. This choice aimed to provide the model with ample learning opportunities to grasp the data dynamics using the specific parameters, enabling a more informed decision-making process. Our experimentation showed that 5 epochs provided an ideal window to comprehend the optimal hyperparameter values.

Subsequently, armed with the parameter configurations that exhibited the best outcomes, we trained the model on the training dataset for 25 epochs. Our observation informed this choice that the models demonstrated robustness and did not exhibit signs of overfitting even when trained for more than 30 epochs. However, mindful of the computational resources required for extensive training, we concluded at the 25-epoch mark. At this point, the model showcased substantial competency and performance, thereby facilitating its subsequent evaluation.

To address the particular research query, we pursued a similar approach. We discerned the optimal configuration for the token matching method to be a batch size of 2, coupled with a learning rate of 5e-05. In contrast, the position-matching method exhibited its utmost effectiveness with a batch size of 2 and a learning rate 3e-05. This meticulous parameter selection process was undertaken to ensure the fine-tuning of our models with the most suitable settings, thereby maximizing their performance potential.

As demonstrated in Figure 5.1, it is evident that the position mapping method consistently outperforms the token matching method across all epochs. Notably, the final F1 scores on the test dataset indicate a value of 0.323 for the token-matching method and 0.407 for the position-matching method. Consequently, for the subsequent experiments, we will exclusively adopt the position-matching approach for token labelling.

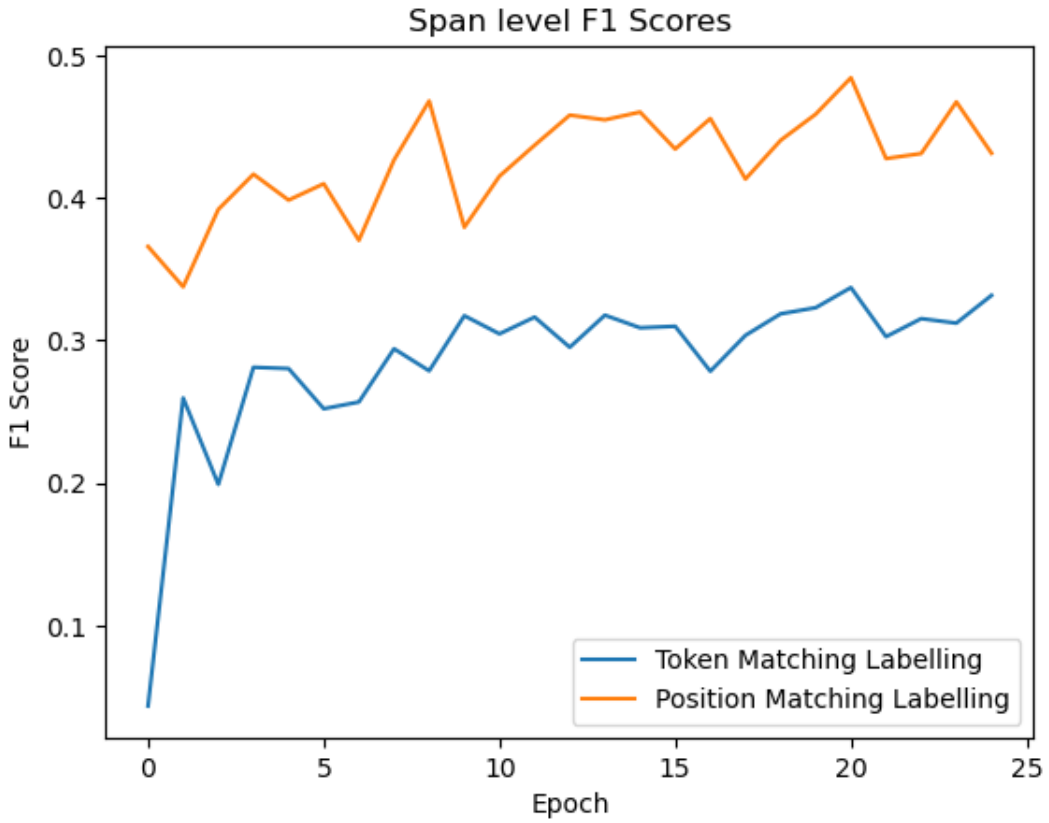

Figure 5.1: The span-level F1 score progression of the token and label matching methods is compared in this analysis. This progression is on the validation set.

Model	Span F1	Precision	Recall
Token Matching	0.323	0.325	0.320
Position Matching	0.407	0.429	0.388

Table 5.1: Results of RQ1 on test set.

5.4 RQ2: How does the performance of specialized models like BioBERT compare to that of a general model like BERT?

Expanding upon the insights garnered from the initial research query, and we opted to continue utilizing the position-matching labelling method for the subsequent phases of experimentation. In response to this inquiry, we trained both the BioBERT and PubMedBERT models, thereby juxtaposing their performance against the position labelling method outcomes previously obtained, which will act as the baseline for this experiment.

The BioBERT model exhibited its optimal performance with a batch size of 8 and a learning rate 5e-05. Conversely, the PubMedBERT model demonstrated its peak performance when trained with a batch size of 2 and a learning rate 5e-05. While Figure 5.2 presents a certain

level of ambiguity regarding the superiority of one model over the other, our evaluation on an independent test dataset clearly showed that PubMedBERT outperformed the other two models.

With a baseline F1 score of 0.407, PubMedBERT achieved 0.448, while BioBERT secured 0.437. While the margins between these scores are slight, they denote a discernible enhancement over the baseline. Consequently, it can be inferred that leveraging contextual information contributes to augmenting the model's overall performance.

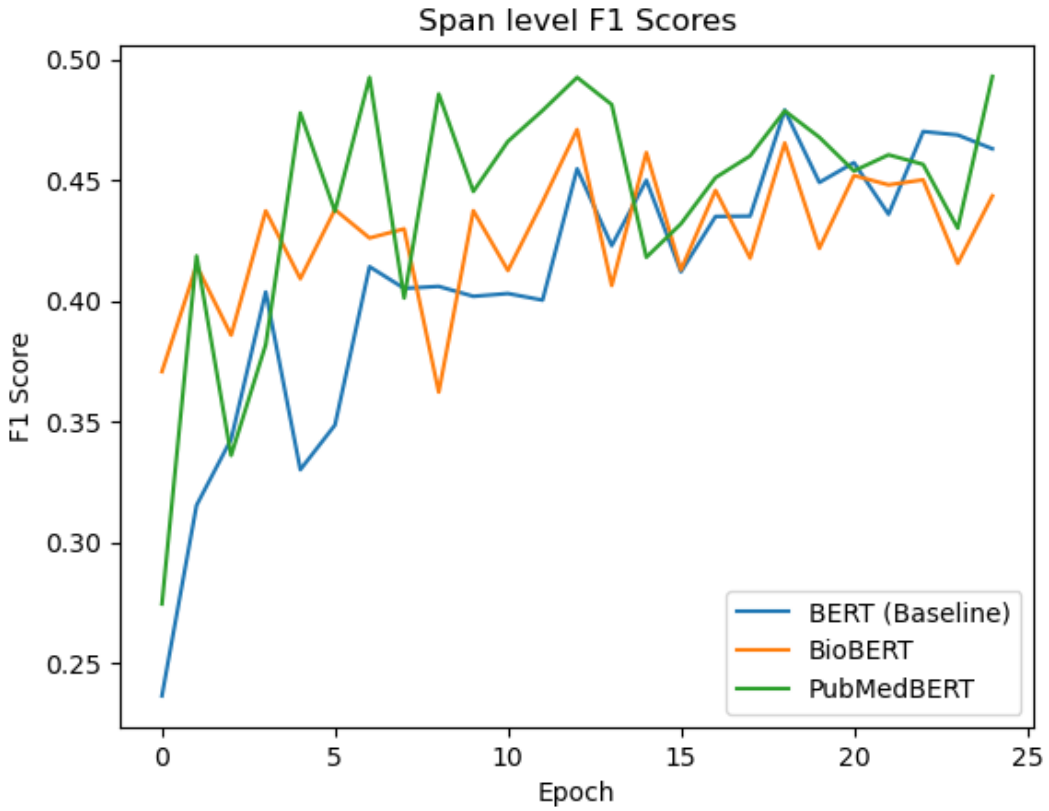

Figure 5.2: The span-level F1 score progression of BERT, BioBERT and PubMedBERT models. This progression is on the validation set.

Model	Span F1	Precision	Recall
Baseline	0.407	0.429	0.388
BioBERT	0.437	0.452	0.422
PubMedBERT	0.447	0.448	0.446

Table 5.2: Results of RQ2 on test set

5.5 RQ3: To what extent does curriculum learning improve the model's performance?

Addressing the final research inquiry, we embarked on a unique approach by initially fine-tuning the PubMed model using BioRED's gene dataset before training on our dataset. This two-stage process began with the BioRED dataset, culminating in a commendable F1 score of 0.736. This achievement was realized by utilizing a batch size of 8 and a learning rate 1e-05.

Subsequently, this fine-tuned model was saved, and the subsequent phase involved training it on our dataset, adhering to the procedures above. The resultant F1 score on the test set was 0.476, achieved with a batch size of 4 and a learning rate of 3e-05. A noteworthy observation is the discernible performance enhancement achieved through curriculum learning when compared against the baseline PubMed model lacking this curriculum-based training.

It is also interesting to note that the BioRED model does not do well on the validation set compared to the PubMedBERT model. However, its performance on the test set is better.

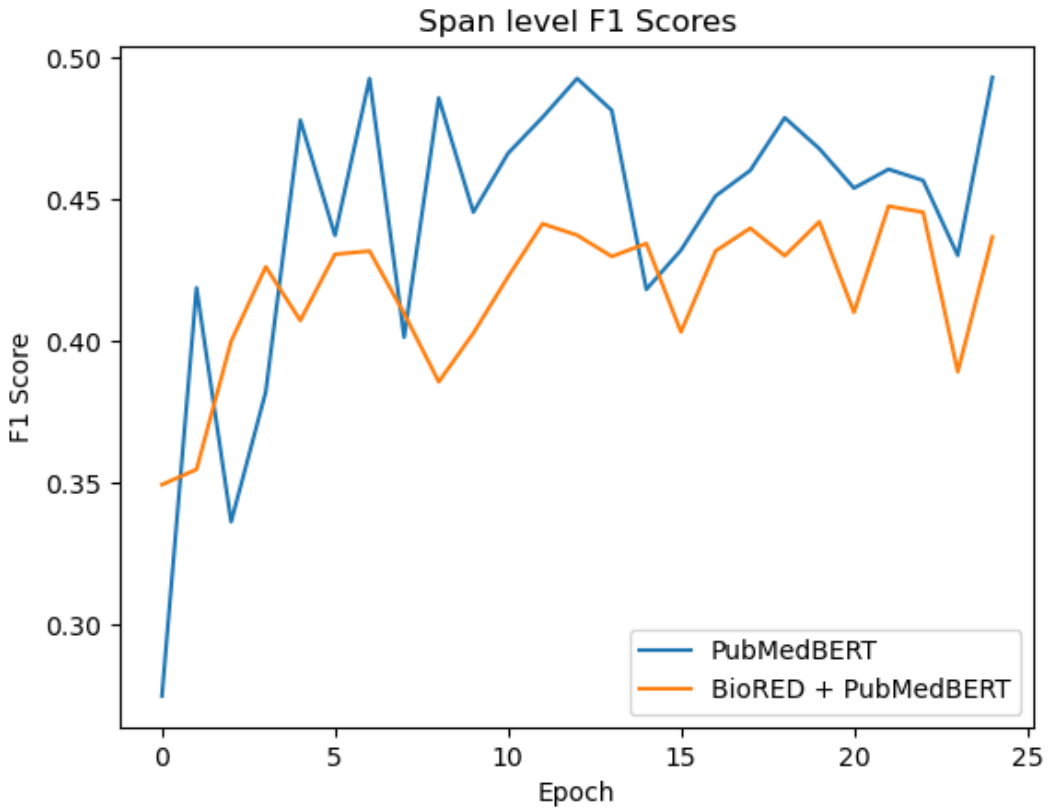

Figure 5.3: The span-level F1 score progression of the PubMedBERT model and BioRED + PubMedBERT model. This progression is on the validation set.

Model	Span F1	Precision	Recall
PubMedBERT	0.447	0.448	0.446
BioRED + PubMedBERT	0.475	0.464	0.487

Table 5.3: Results of RQ3 on test set

During the evaluation, we encountered some interesting findings. We achieved an increase of 48% through incorporating curriculum learning, as compared to the unmodified BERT model's performance on the Token Matching Labelling method, which is the baseline for the project. This outcome underscores the efficacy of curriculum learning in elevating model performance.

Model	Span F1	Precision	Recall
BERT (Token)	0.323	0.325	0.320
BERT (Position)	0.407	0.429	0.388
BioBERT	0.437	0.452	0.422
PubMedBERT	0.447	0.448	0.446
BioRED + PubMedBERT	0.475	0.464	0.487

Table 5.4: Results of all models on test set

Chapter 6: Conclusion

6.1 Achievements

In line with our overarching research pursuits, our project aimed to construct a system that could assist biomedical researchers in extracting essential metadata for their investigations. This endeavour led us to curate a dedicated database for training models in genotype recognition within the biomedical realm. Through this resourceful dataset, we accomplished three distinct research objectives. We determined that employing the positional mapping method for text labelling in BioNER tasks yielded superior results than token matching. We also identified that the PubMed model outperformed the rest by evaluating different models on our dataset. Lastly, our implementation of curriculum learning offered valuable insights, indicating a discernible enhancement in model performance. We achieved a 48% improvement in performance compared to the unmodified BERT model. In conclusion, our project contributes a roadmap for metadata extraction from unannotated samples and underscores the potential of curriculum learning for advancing biomedical NLP endeavours. These achievements collectively position our work as a substantial stride toward more efficient and accurate biomedical entity recognition.

6.2 Limitations and Future work

While this project has provided valuable insights and contributions to biomedical entity recognition, several limitations should be acknowledged.

1. The hyperparameter range could have been expanded to improve the model's performance further. However, due to resource constraints, the search for optimal hyperparameters was limited. A broader exploration of hyperparameter values could have yielded even better results.
2. The dataset used for training and evaluation was constrained by the limit of 10,000 entries imposed by the API. Access to a larger and more diverse dataset could have enhanced the model's ability to generalize to different scenarios, potentially resulting in improved performance.
3. The curriculum learning approach was implemented in a basic form due to time limitations. Exploring a more in-depth curriculum learning strategy could have led to even greater performance improvements in the model, offering a richer understanding of the potential benefits of this technique.
4. A more comprehensive evaluation could have involved using the trained model to make predictions on text without explicitly marked genotypes. This would have provided insights into the model's robustness and generalization capabilities beyond the specific labelled entities.

Additionally, Exploring the horizons of future work within this project unveils compelling avenues for further advancement. One of the pivotal challenges encountered in BioNER is the inconsistency in naming conventions present within the biomedical texts. This issue

poses an exciting research opportunity to develop methodologies capable of recognizing and reconciling diverse variations of the same entity within the text. Such an endeavour would significantly amplify the capacity of NER models to identify entities across a broader spectrum of contexts.

Appendix A: Code

The code for this project is hosted on GitHub: https://github.com/dhruvkp090/bioNER_dissertation

Bibliography

[1] Bio.entrez package — biopython 1.75 documentation.

[2] Epigenetic treatment of behavioral and physiologic... (ID 755255) - BioProject - NCBI.

[3] Hepatocytes direct the formation of a pro-metastat... (ID 431170) - BioProject - NCBI.

[4] How do your biomedical named entity recognition models generalize to novel entities? 10:31513–31523.

[5] KPC 3 - BioSample - NCBI.

[6] Metadata archives.

[7] T. Barrett, K. Clark, R. Gevorgyan, V. Gorelenkov, E. Gribov, I. Karsch-Mizrachi, M. Kimelman, K. D. Pruitt, S. Resenchuk, T. Tatusova, E. Yaschenko, and J. Ostell. BioProject and BioSample databases at NCBI: facilitating capture and organization of metadata. 40:D57–D63.

[8] National Research Council (US) Board on Biology, Robert Pool, and Joan Esnayra. Barriers to the use of databases. In *Bioinformatics: Converting Data to Knowledge: Workshop Summary*. National Academies Press (US).

[9] Roland Brian Büchter, Alina Weise, and Dawid Pieper. Development, testing and use of data extraction forms in systematic reviews: a review of methodological guidance. 20:259.

[10] Maria Carmela Cariello, Alessandro Lenci, and Ruslan Mitkov. A comparison between named entity recognition models in the biomedical domain. In *Proceedings of the Translation and Interpreting Technology Online Conference*, pages 76–84. INCOMA Ltd.

[11] Gamal Crichton, Sampo Pyysalo, Billy Chiu, and Anna Korhonen. A neural network multi-task learning approach to biomedical named entity recognition. 18(1):368.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding.

[13] John M Giorgi and Gary D Bader. Transfer learning for biomedical named entity recognition with neural networks. 34(23):4087–4094.

[14] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. 3(1):1–23.

[15] Ángel Gálvez-Merchán, Kyung Hoi (Joseph) Min, Lior Pachter, and A Sina Booeshaghi. Metadata retrieval from sequence databases with ffq. 39(1):btac667.

[16] Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. Deep learning with word embeddings improves biomedical named entity recognition. 33(14):i37–i48.

[17] Cody E. Hinchliff and Stephen Andrew Smith. Some limitations of public sequence data for phylogenetic inference (in plants). 9(7):e98986.

[18] Tim Hulsen, Saumya S. Jamuar, Alan R. Moody, Jason H. Karnes, Orsolya Varga, Stine Hedensted, Roberto Spreafico, David A. Hafler, and Eoin F. McKinney. From big data to precision medicine. 6:34.

[19] Qiao Jin, Bhuwan Dhingra, William W. Cohen, and Xinghua Lu. Probing biomedical embeddings from language models.

[20] Siddhartha R. Jonnalagadda, Pawan Goyal, and Mark D. Huffman. Automating data extraction in systematic reviews: a systematic review. 4:78.

[21] Donghyeon Kim, Jinhyuk Lee, Chan Ho So, Hwisang Jeon, Minbyul Jeong, Yonghwa Choi, Wonjin Yoon, Mujeen Sung, and Jaewoo Kang. A neural named entity recognition and multi-type normalization tool for biomedical text mining. 7:73729–73740. Conference Name: IEEE Access.

[22] Adam Klie, Brian Y Tsui, Shamim Mollah, Dylan Skola, Michelle Dow, Chun-Nan Hsu, and Hannah Carter. Increasing metadata coverage of SRA BioSample entries using deep learning–based named entity recognition. 2021:baab021.

[23] Seth Kulick, Ann Bies, Mark Liberman, Mark Mandel, Ryan McDonald, Martha Palmer, Andrew Schein, Lyle Ungar, Scott Winters, and Pete White. Integrated annotation for biomedical information extraction. In *HLT-NAACL 2004 Workshop: Linking Biological Literature, Ontologies and Databases*, pages 61–68. Association for Computational Linguistics.

[24] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. 36(4):1234–1240.

[25] Ki-Joong Lee, Young-Sook Hwang, Seonho Kim, and Hae-Chang Rim. Biomedical named entity recognition using two-phase model based on SVMs. 37(6):436–447.

[26] R. Leinonen, H. Sugawara, M. Shumway, and on behalf of the International Nucleotide Sequence Database Collaboration. The sequence read archive. 39:D19–D21.

[27] Angli Liu, Jingfei Du, and Veselin Stoyanov. Knowledge-augmented language model and its application to unsupervised named-entity recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1142–1150. Association for Computational Linguistics.

[28] Hongfang Liu, Zhang-Zhi Hu, Manabu Torii, Cathy Wu, and Carol Friedman. Quantitative assessment of dictionary-based protein named entity tagging. 13(5):497–507.

[29] Shengyu Liu, Buzhou Tang, Qingcai Chen, and Xiaolong Wang. Drug name recognition: Approaches and resources. 6(4):790–810. Number: 4 Publisher: Multidisciplinary Digital Publishing Institute.

[30] Supratim Mukherjee, Dimitri Stamatis, Cindy Tianqing Li, Galina Ovchinnikova, Jon Bertsch, Jagadish Chandrabose Sundaramurthi, Mahathi Kandimalla, Paul A Nicolopoulos, Alessandro Favognano, I-Min A Chen, Nikos C Kyrpides, and T B K Reddy. Twenty-five years of genomes OnLine database (GOLD): data updates and new features in v.9. 51:D957–D963.

[31] Yifan Peng, Shankai Yan, and Zhiyong Lu. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65. Association for Computational Linguistics.

[32] Eric Sayers. A general introduction to the e-utilities. In *Entrez Programming Utilities Help [Internet]*. National Center for Biotechnology Information (US).

[33] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pages 107–110. COLING.

[34] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey.

[35] NCBI Staff. The entire corpus of the sequence read archive (SRA) now live on two cloud platforms!

[36] Yuanhe Tian, Wang Shen, Yan Song, Fei Xia, Min He, and Kenli Li. Improving biomedical named entity recognition with syntactic information. 21(1):539.

[37] Hannes Ulrich, Ann-Kristin Kock-Schoppenhauer, Noemi Deppenwiese, Robert Gött, Jori Kern, Martin Lablans, Raphael W Majeed, Mark R Stöhr, Jürgen Stausberg, Julian Varghese, Martin Dugas, and Josef Ingenerf. Understanding the nature of metadata: Systematic review. 24(1):e25440.

[38] Zihan Wang, Jingbo Shang, Liyuan Liu, Lihao Lu, Jiacheng Liu, and Jiawei Han. CrossWeigh: Training named entity tagger from imperfect annotations.