Integrating Vector Stores and Information Retrieval for Analysis Literature Data on *Mycobacterium tuberculosis*: experience feedback

Christophe Guyeux
Institut FEMTO-ST, UMR 6174 CNRS
Université de Franche-Comté
Belfort, France
christophe.guyeux@univ-fcomte.fr

***Abstract*—Tuberculosis, caused by Mycobacterium tuberculosis, remains a significant public health issue globally, exacerbated by the growing prevalence of multidrug-resistant strains and the lack of new therapeutic options. Despite extensive research efforts, there remains a critical need to effectively harness the vast amount of available data to drive new insights and treatment strategies. Currently, over 200,000 genomes of Mycobacterium tuberculosis are publicly available, and there are more than 100,000 scientific articles on PubMed concerning this bacterium. The potential for breakthroughs is immense if we can systematically study these two corpora in conjunction. However, the scale of these datasets necessitates the development of new tools based on big data analytics, artificial intelligence, and large language models (LLMs). This paper presents a study on the application of information retrieval (IR) techniques for performing Retrieval-Augmented Generation (RAG) to facilitate access to current knowledge about Mycobacterium tuberculosis. The approach integrates IR and response generation to provide relevant and contextually appropriate answers to user queries, leveraging scientific documents and genomic data.**

***Index Terms*—Mycobacterium tuberculosis; Vector Stores; Information Retrieval; Retrieval-Augmented Generation; Genomic Analysis**

I. Introduction

Tuberculosis (TB), caused by Mycobacterium tuberculosis, remains a formidable global health challenge despite decades of intensive research and public health efforts [29]. The disease continues to afflict millions worldwide, with the World Health Organization reporting approximately 10 million new cases and 1.5 million deaths annually [25]. Alarmingly, the rise of multidrug-resistant (MDR) and extensively drug-resistant (XDR) strains of M. tuberculosis has exacerbated the situation, rendering standard treatment regimens ineffective and complicating disease management [19]. This worrying evolution underscores the urgent need for novel strategies to understand, prevent, and treat TB.

The advent of high-throughput genomic sequencing and the exponential growth of scientific publications have resulted in an unprecedented accumulation of data related to M. tuberculosis [19]. Over 200,000 bacterial genomes are publicly available [15], providing a rich resource for genomic analysis and epidemiological studies. Additionally, more than 100,000 scientific articles on PubMed discuss various aspects of this pathogen. While this wealth of data holds immense potential for breakthroughs in TB research, it also presents significant challenges. The sheer volume and complexity make it difficult for researchers to stay current, extract relevant information, and synthesize insights across disparate data sources.

Emerging artificial intelligence (AI) technologies, particularly large language models (LLMs), offer promising solutions to navigate and interpret vast datasets [20]. Techniques such as Information Retrieval (IR) and Retrieval-Augmented Generation (RAG, see Figure 1) enable the extraction of pertinent information and the generation of contextually appropriate responses to user queries [20]. These tools can potentially revolutionize how researchers access and integrate knowledge from genomic data and scientific literature, facilitating more efficient discovery and innovation in TB research.

However, leveraging these AI-based tools effectively is not without challenges. A purely semantic IR approach may overlook novel findings that are documented in only a few recent publications, such as the discovery of a new lineage like lineage 10 of *M. tuberculosis* [12]. Consequently, if a researcher queries the number of known lineages, the system might erroneously report a maximum of nine, ignoring the latest developments. Additionally, scientific knowledge about M. tuberculosis evolves over time due to advancements in techniques – from phenotypic characterizations to spoligotyping, MIRU-VNTR, and increasingly precise SNP-based definitions as more genomes are sequenced [1], [4], [14], [28]. Older, superseded information may persist in the literature, leading to potential inaccuracies if not properly contextualized. Moreover, the credibility of sources varies; findings from highly cited, reputable authors may carry more weight than those from less established researchers, especially when presenting surprising or groundbreaking results.

The goal of this paper is to address these complexities by sharing our experiences and proposing strategies to enhance the relevance and accuracy of AI-assisted tools in TB research. We aim to explore methods for improving IR and RAG systems to account for the temporal evolution of knowledge, the significance of author reputation, and the need to highlight

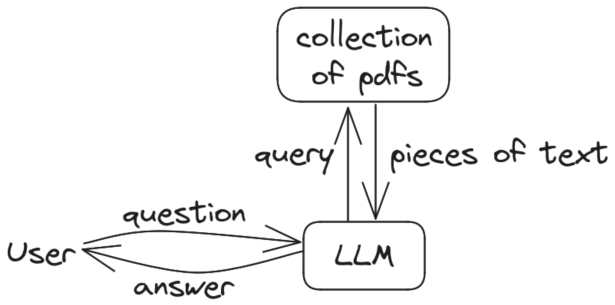

Fig. 1. Retrieval-Augmented Generation (RAG): a language model (LLM) queries a collection of PDFs to retrieve relevant text snippets. The LLM uses these snippets to generate a more accurate answer to the user's question by grounding its response in external information.

recent, yet crucial, scientific discoveries. By tackling these challenges, we aspire to provide valuable insights and practical solutions that will assist researchers in effectively harnessing the vast genomic and literature data on Mycobacterium tuberculosis, ultimately contributing to more informed and impactful advancements in the fight against tuberculosis.

The remainder of this article is as follows. The next section is a state of the art related to information retrieval and its application in the bioinformatics field. In Section III, the final objecives and its associated scientific challenges are introduced. Avenues for solutions are then proposed in Section IV. This article ends by a conclusion section, in which the contributions are summarized and intended future work is outlined.

II. State of the Art

A. Information Retrieval

Information Retrieval (IR) has long been a pivotal field in computer science, dedicated to the organization and retrieval of information from large datasets, such as collections of research articles. Early IR systems primarily relied on keyword-based methods, utilizing statistical techniques like Term Frequency-Inverse Document Frequency (TF-IDF) to rank documents based on the occurrence of query terms [27]. While effective to an extent, these models faced limitations in understanding the semantic relationships between words, often struggling with issues like synonymy and polysemy.

The introduction of Latent Semantic Analysis (LSA) marked a significant advancement by capturing underlying semantic structures in text data [8]. LSA reduced the dimensionality of term-document matrices, uncovering latent relationships between terms and documents. However, these models still lacked the ability to handle context dynamically. The emergence of word embeddings, such as Word2Vec [23] and GloVe [26], revolutionized IR by representing words in continuous vector spaces where semantic similarities are encoded geometrically. This advancement allowed IR systems to understand context and semantic nuances better, improving the retrieval of relevant documents.

The advent of transformer architectures [30] brought about a paradigm shift in natural language processing (NLP) and, by extension, IR. Transformers leveraged self-attention mechanisms to capture long-range dependencies in text, enabling models to understand context more effectively than previous recurrent architectures. Building on transformers, large language models (LLMs) like BERT [9] provided deep bidirectional representations of text, significantly enhancing the performance of IR systems, especially in handling complex queries typical in academic research. Furthermore, the integration of IR with generative models in approaches like Retrieval-Augmented Generation (RAG [21]) has enabled the development of systems that can retrieve relevant information and generate contextually appropriate responses, offering substantial benefits for researchers navigating vast amounts of literature.

B. Large Language Models

The advent of Large Language Models (LLMs) has significantly impacted medical research, offering advanced tools for data analysis, knowledge extraction, and decision support. In the field of microbiology, LLMs have been instrumental in processing and interpreting vast amounts of textual data, such as scientific literature, clinical reports, and genomic information. For instance, models like BioBERT [20] and ClinicalBERT [2] have been specifically fine-tuned for biomedical text mining, enhancing tasks such as named entity recognition, relation extraction, and question answering within medical texts.

In microbiology, LLMs have facilitated the rapid identification and characterization of pathogens by extracting relevant information from unstructured data sources. They have been used to predict microbial gene functions and interactions by analyzing scientific publications and genomic datasets [16]. Furthermore, LLMs contribute to antimicrobial resistance research by identifying patterns and correlations in large-scale genomic and clinical data, aiding in the prediction of resistance mechanisms [3]. These applications not only accelerate the pace of discovery but also improve the accuracy of microbiological analyses.

The integration of LLMs into microbiological research workflows holds great promise for accelerating discoveries and improving public health outcomes. Ongoing research is exploring the use of LLMs in conjunction with other AI techniques, such as deep learning models for protein structure prediction, exemplified by AlphaFold [17], and in the development of intelligent systems for real-time pathogen surveillance and outbreak detection [7]. As these models continue to evolve, their applications are expected to expand, offering more sophisticated tools for understanding complex microbial behaviors and interactions.

III. Final Objective and Scientific Challenges

For an LLM-based tool to be genuinely useful to a researcher working in a highly specific and specialized field (e.g., the phylogeny of *M. tuberculosis*), the foundational knowledge of an LLM is vastly insufficient [13]. Even LLMs

specifically trained on medical corpora remain far too generalist, and only RAG approaches utilizing the most comprehensive corpus on *M. tuberculosis* can lead to an LLM-based copilot truly beneficial for the researcher.

Building such a corpus is a challenge in itself, as it must encompass the entire scientific literature on this bacterium – a corpus that is difficult to assemble – and also incorporate newly published articles daily. Moreover, it should not focus solely on the species level but adopt a broader perspective: encompassing literature on the entire *Mycobacterium* genus (for questions about the origin of the species or its distinction from close relatives), as well as relevant insights from microbiology, bioinformatics, statistics, and even population history (e.g., migration patterns). This initial difficulty is indeed a major challenge but lies outside the scope of this paper, which assumes that such a knowledge base has already been established (cf. [13]).

The next challenge, given a researcher's query, is to extract scientific text snippets with the highest likelihood of containing the answer. Currently, the effectiveness of the tool hinges on the precision of this information retrieval step: in practice, today's best "generalist" LLMs are fully capable of providing highly relevant answers that meet a researcher's expectations, provided they are given the right text excerpts. However – and this is a key point – while it is clear that older keyword-based or TF-IDF approaches are inadequate for this task, it is equally true that a purely semantic approach is insufficient. The solution lies in an iterative hybrid approach, as presented in the next section.

Indeed, while semantic search in vector stores has significantly improved information retrieval, relying solely on semantic similarity can be insufficient for specialized research domains like microbiology. A hybrid search approach that combines semantic understanding with explicit identification of specific entities such as strain identifiers or single nucleotide polymorphisms (SNPs) is essential for effective Retrieval-Augmented Generation (RAG) in this context.

For instance, consider the study of the virulence of *Mycobacterium tuberculosis* lineage 4.6.2 in Nigeria. This lineage was previously known as the "Cameroon" lineage, a term that refers not to the country but to a specific bacterial lineage initially identified there [11]. Over time, taxonomic revisions by researchers such as Freschi and Napier have updated this nomenclature [10], [24]. The dynamic nature of lineage definitions means that purely semantic searches might miss relevant literature if they do not account for historical naming conventions and taxonomic changes. A hybrid approach that includes explicit lineage identifiers ensures comprehensive retrieval of all pertinent information, regardless of nomenclatural evolution.

Furthermore, constructing a database of bacterial strains linked to their antibiotic resistances – a critical resource for developing machine learning models to predict resistance based on genetic mutations – requires precise extraction of strain-specific data [22]. Antibiotic resistances are often reported using various abbreviations and terminologies across different studies. Identifying explicit mentions of strains and standardizing the extraction of resistance profiles necessitates more than semantic similarity; it requires pattern recognition and entity extraction capabilities that can handle domain-specific jargon and notation.

Additionally, valuable information resides outside traditional scientific articles. Repositories like the NCBI database provide metadata for submitted strains, including isolation dates and geographic locations [5]. Databases such as MycoBrowser offer detailed gene-specific information [18]. Integrating these heterogeneous data sources into the retrieval process demands a hybrid approach that can navigate structured databases and unstructured text, combining semantic understanding with precise entity recognition.

These examples highlight the necessity of a hybrid information retrieval strategy that supplements semantic search with explicit entity identification and handling of domain-specific knowledge. Such an approach enhances the effectiveness of RAG systems, enabling researchers to access comprehensive and accurate information crucial for advancing microbiological research, as summarized in Table I.

IV. Avenues for solutions

A. Filling up the Vectorstore

After explaining why a hybrid approach to information retrieval is necessary, the next question is how to implement it. First and foremost, the choice of vectorstore is crucial. There are many options available, but the one selected must be capable of storing not only the text and its embedding but also additional information such as the identifiers of cited strains, the list of authors, cited drugs, and more. This vectorstore must support semantic queries with filters (e.g., texts mentioning a specific drug and a particular country). To meet these needs, we propose using Milvus [31], which supports all these requirements and additionally allows for querying across multiple vector columns simultaneously.

Populating this vectorstore does not present any specific challenges, but it does require implementing customized post-processing for each chunk. After parsing each PDF and segmenting it into paragraphs (for instance), the additional columns can be filled using ad hoc methods. These methods may include substring detection (for antibiotics), regular expressions (e.g., strain identifiers in the format of E, S, or D followed by RR and an integer), or similarity metrics like Levenshtein distance based on a predefined list of locations (countries, cities, etc.).

Bibliographic information (title, authors, journal name, etc.) is also crucial to retrieve – first, to provide the user with the sources used to answer their query, and second, to filter the returned excerpts before providing them to the LLM, as we will see below. However, retrieving the journal name from the PDF can be challenging, primarily due to the variations in style and layout across different scholarly journals. Identifying author names is similarly difficult, as they may contain special characters that complicate OCR processing. To address this challenge, the approach involves first consolidating the title (by

TABLE I
COMPARISON OF INFORMATION RETRIEVAL APPROACHES

Approach	Pros	Cons
Embeddings/Semantic Search Alone	• Captures contextual and semantic meaning of text. • Handles synonymy and polysemy effectively. • Useful for general topic exploration.	• May miss specific domain entities like strain identifiers or SNPs. • Struggles with evolving terminology and nomenclature changes. • Recent, less frequent findings might be overlooked.
Keyword-Based Search Alone	• Precise retrieval of documents containing specific terms or identifiers. • Effective for extracting explicit mentions of strains or resistance profiles. • Simple implementation and fast execution.	• Lacks understanding of context and semantics. • Unable to handle synonyms or related concepts. • High risk of missing relevant documents due to terminology variations.
Hybrid Approach	• Combines semantic understanding with precise entity recognition. • Accounts for evolving terminology and nomenclature changes. • Enhances retrieval of recent and significant findings. • Integrates diverse data sources (articles, databases, metadata).	• Increased complexity in system design and implementation. • Requires more computational resources. • Needs careful tuning to balance semantic and keyword components.

inspecting metadata, automatically searching for the closest match on PubMed, etc.), and then automatically retrieving the associated BibTeX entry, for instance, using the CrossRef API [6].

Other types of information may be more challenging to retrieve. Taking the case of lineages as an example, there has been a shift from an early taxonomy based on names (which can be confused with locations, such as Ghana, Haarlem, Uganda, Cameroon, West African, or with mathematical symbols, like X-type and X2 lineages) to a numerical nomenclature (e.g., L4.1, L4.3.2...). In the first case, using an LLM with a custom prompt can help determine, based on context, whether these terms refer to lineages, locations, or something else whenever one of these older terms is detected. In the second case, the previously mentioned issue is that lineage schemes evolve over time. Currently, there is no definitive taxonomy (new lineages have been discovered recently), and there are compatibility breaks between lineages. A different prompt could be used to determine the taxonomic scheme based on the full article content, followed by an ad hoc script to convert it into an absolute taxonomy, thus removing ambiguity. The complex case of articles using multiple taxonomic schemes, for comparison purposes, would also need to be handled. Properly managing lineage information is, as we can see, complex to implement but feasible, and absolutely essential to avoid returning incorrect information.

B. Filtering the Retrieved Information

When the vectorstore only contains a single column of embeddings, information retrieval is quite limited: it simply involves finding texts whose embeddings are close (e.g., using cosine similarity or dot product) to the embedding of the user's query, potentially adding some diversity. Having multiple columns in the vectorstore allows for querying in various ways (e.g., filtering for texts that mention a specific country or originate from recent articles), then extracting a sample of texts from each query, which provides the LLM with a potentially richer and more varied set of chunks. To further enhance the number and relevance of the excerpts passed to the LLM, prompt engineering can be applied by asking an LLM to reformulate the query, generalize it, or split it into more basic sub-queries, etc. Such approaches can be scaled up, given the reduction in response time and cost when using recent LLMs.

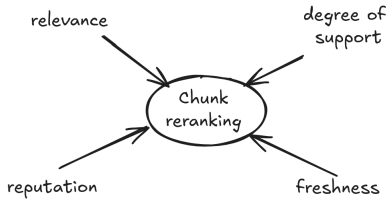

Fig. 2. Reranking approaches to consider in the Information Retrieval process.

However, despite the increase in allowed context sizes in these modern LLMs, various experiments have shown that providing a context of reasonable length, with few low-quality texts in terms of relevance to the initial query, yields better results. This is why a second stage of ranking and filtering the initial results is typically implemented.

Effective reranking must be able to consider, based on a fact contained within the chunk: the freshness of the fact (is it recent information?), its relevance (in relation to the query), its degree of support (other chunks corroborate it, with few contradictory sources), and the reputation of the "bearer" of the fact (a well-established author, a reputable journal, or a highly cited article), see Figure 2.

The recency of information can be determined simply by consulting the publication date extracted from the BibTeX entry. The relevance of an excerpt for answering the initial query can be assessed by an LLM with a structured output, using a prompt like: "On a scale of 1 to 5, to what extent does the following excerpt address the query...". A similar, though more challenging, approach can be used to gauge the degree of support: an initial prompt can list conflicting trends within the returned excerpts, followed by a second prompt that classifies each excerpt according to its trend.

The reputation associated with the text's producer can be based on various metrics retrievable online via APIs: the average H-index of the authors, the impact factor of the journal, the article's citations, etc. For the first and last authors listed, it's also possible to use LangChain's PubMed retriever to gather abstracts of the authors' recent publications and see if they have been actively producing work on topics related to the initial query. This can be executed with agents (LLMs equipped with "tools"), as shown in Figure 3. While the number of requests associated with such approaches currently appears prohibitive – both in terms of execution time and cost – the ongoing improvements in LLMs will make this feasible very soon.

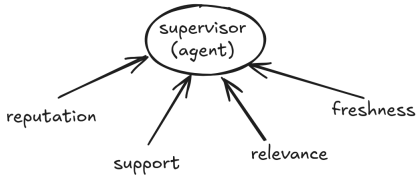

Fig. 3. Agentic approach for Information Retrieval.

V. Conclusion

The challenges and strategies associated with using vector stores and information retrieval (IR) for analyzing literature data on Mycobacterium tuberculosis (MTB) have been explored in this article. Given the vast volume of MTB-related genomic and scientific data, traditional keyword-based IR is insufficient for high specificity needs in such research. We argued for a hybrid retrieval approach that combines semantic search with explicit entity recognition (e.g., strain identifiers or SNPs), essential for effectively supporting retrieval-augmented generation (RAG) in microbiological research. By leveraging big data analytics, AI, and large language models (LLMs), this study aimed to enable more contextually relevant responses to user queries, aiding researchers in quickly navigating a complex and evolving knowledge landscape.

Future work in this domain could focus on enhancing the precision and adaptability of the proposed information retrieval system. One promising direction is the development of dynamic taxonomic mapping that adapts in real-time to updates in MTB lineages and nomenclature changes. This approach would involve implementing a lineage tracking mechanism that continuously integrates the latest taxonomic insights, ensuring the retrieval system remains relevant and comprehensive despite the evolving nature of scientific understanding in microbiology. Moreover, as LLM capabilities advance, incorporating multi-step querying agents that can reformulate complex queries, identify implicit connections across articles, and weigh evidence from disparate sources will be essential. These agents could refine search outcomes by iteratively interacting with the vectorstore, aiming to retrieve the most contextually relevant information.

In addition, expanding the hybrid retrieval framework to include external structured datasets – such as antimicrobial resistance databases or real-time epidemiological data – could provide a richer, multidimensional perspective on MTB research. Integrating these diverse data sources would not only enhance the quality of RAG-based outputs but also enable predictive analytics, potentially allowing researchers to forecast trends in drug resistance or identify emerging MTB strains. Finally, future work could address the computational efficiency of such advanced systems, with a particular focus on optimizing response times and reducing costs, as computational demands continue to increase with the scale and complexity of integrated data.

References

[1] Céline Allix-Béguec, Dag Harmsen, Thomas Weniger, Philip Supply, and Stefan Niemann. Evaluation and strategy for use of miru-vntr plus spoligotyping for molecular epidemiology of tuberculosis. *Journal of Clinical Microbiology*, 46(8):2648–2656, 2008.

[2] Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. Publicly available clinical bert embeddings. *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78, 2019.

[3] German Arango-Argoty, Emily Garner, Amy Pruden, Lenwood S Heath, Peter Vikesland, and Liqing Zhang. Deeparg: a deep learning approach for predicting antibiotic resistance genes from metagenomic data. *Microbiome*, 6(1):1–15, 2018.

[4] Francesc Coll, Ruth McNerney, João A Guerra-Assunção, Judith R Glynn, Joana Perdigão, Miguel Viveiros, Isabel Portugal, Arnab Pain, Nick Martin, and Taane G Clark. Whole-genome sequencing for molecular epidemiology of mycobacterium tuberculosis. *New England Journal of Medicine*, 369(20):1901–1910, 2013.

[5] NCBI Resource Coordinators. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 46(D1):D8–D13, 2018.

[6] CrossRef Organization. Crossref. Accessed: 2024-10-31, 2024. https://www.crossref.org.

[7] Ashlynn R Daughton, Nicholas Generous, Reid Priedhorsky, et al. An open-source, customizable, and scalable platform for real-time epidemic tracking and forecasting. *Influenza and Other Respiratory Viruses*, 14(2):212–217, 2020.

[8] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990.

[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186, 2019.

[10] Luca Freschi, Rafael Vargas, Ahsan Husain, Sarah Kamal, Alena Skrahina, Saima Tahseen, Natalia Kurepina, Maxime Seghers, and et al. Population structure, biogeography and transmissibility of mycobacterium tuberculosis. *Nature Communications*, 12(1):1–14, 2021.

[11] Sebastien Gagneux, Kathryn DeRiemer, Tinh Van, Midori Kato-Maeda, Bouke C de Jong, Sathish Narayanan, Mark Nicol, Stefan Niemann, Kristin Kremer, Maria C Gutierrez, et al. Variable host-pathogen compatibility in mycobacterium tuberculosis. *Proceedings of the National Academy of Sciences*, 103(8):2869–2873, 2006.

[12] Christophe Guyeux, Gaetan Senelle, Adrien Le Meur, Philip Supply, Cyril Gaudin, Jody E Phelan, Taane G Clark, Leen Rigouts, Bouke de Jong, Christophe Sola, et al. Newly identified mycobacterium africanum lineage 10, central africa. *Emerging Infectious Diseases*, 30(3):560, 2024.

[13] Christophe Guyeux, Christophe Sola, and David Laiymani. Leveraging llm-powered systems to accelerate mycobacterium tuberculosis research step one: From documents to the vectorstore. In *The 10th International Conference on machine Learning, Optimization and Data science - LOD 2024*, pages ***–***, September 2024.

[14] Christophe Guyeux, Christophe Sola, and Guislaine Refrégier. Crisprbuilder-tb: "crispr-builder for tuberculosis". exhaustive reconstruction of the crispr locus in mycobacterium tuberculosis complex using sra. *PLOS Computational Biology*, 17(3), March 2021.

[15] Sarah M Gygli, Sonia Borrell, Andrej Trauner, and Sebastien Gagneux. Practical guidance for standardizing tuberculosis biomarker studies: value of replication and critical analysis of pitfalls. *Journal of Infectious Diseases*, 220(Suppl 3):S176–S185, 2019.

[16] Xiang Ji, Li Yao, and Rui Jiang. Emerging trends in computational microbiology. *Genome Biology*, 22(1):1–23, 2021.

[17] John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[18] Alexandra Kapopoulou, Jun-Rong Lew, and Stewart T Cole. The mycobrowser portal: a comprehensive and manually annotated resource for mycobacterial genomes. *Tuberculosis*, 91(1):8–13, 2011.

[19] Christoph Lange, Keertan Dheda, Dumitru Chesov, Anna Mandalakas, Zarir Udwadia, and Charles R Horsburgh. Management of drug-resistant tuberculosis. *The Lancet*, 394(10202):953–966, 2019.

[20] 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. *Bioinformatics*, 36(4):1234–1240, 2020.

[21] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474, 2020.

[22] Conor J Meehan, Pablo Moris, Thomas A Kohl, Rafael Peña-Miller, Dustin L Dolinger, Felix Kitavi, Jaroslaw Dziadek, Jesse Vornhagen, and et al. The relationship between transmission time and clustering methods in mycobacterium tuberculosis epidemiology. *EBioMedicine*, 37:410–416, 2018.

[23] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[24] Gabriel Napier, Susana Campino, Yemisirach Merid, Mulugeta Abebe, Abraham Aseffa, Martin L Hibberd, Arnab Pain, and Taane G Clark. Evolutionary history provides insights into the emergence and adaptation of mycobacterium tuberculosis beijing lineage in china. *Scientific Reports*, 10(1):1–11, 2020.

[25] World Health Organization. Global tuberculosis report 2022, 2022. Licence: CC BY-NC-SA 3.0 IGO.

[26] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[27] Gerard Salton and Michael J McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, Inc., 1988.

[28] Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, and Christophe Sola. Investigating the diversity of tuberculosis spoligotypes with dimensionality reduction and graph theory. *Genes*, 13(12), December 2022. Special Issue Selected Papers from the 9th International Work-Conference on Bioinformatics and Biomedical Engineering (IWBBIO 2022).

[29] Christophe Sola, Igor Mokrousov, Muhammed Rabiu Sahal, Kevin La, Gaetan Senelle, Christophe Guyeux, Guislaine Refrégier, and Emmanuelle Cambau. Chapter 27 - evolution, phylogenetics, and phylogeography of mycobacterium tuberculosis complex. 2024.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[31] Jianguo Wang, Yifei Li, Xupeng Li, et al. Milvus: A purpose-built vector data management system. *Proceedings of the 46th International Conference on Very Large Data Bases (VLDB)*, 2020. https://milvus.io.