Advancing Scientific Knowledge Representation: Standardisation and Integration in Tolerogenic Therapies

By

Ayesha Sahar

Thesis

submitted in fulfilment of the requirements for the Degree of

Doctor of Philosophy

School of Computing Science
Newcastle University
2024

Statement of Originality

The thesis contains no material which has been accepted for the award of any other degree or diploma in any university or other tertiary institution and, to the best of my knowledge and belief, contains no material previously published or written by another person, except where due reference has been made in the text. I give consent to the final version of my thesis being made available worldwide when deposited in the University's Digital Repository**, subject to the provisions of the Copyright Act 1968.

**Unless an Embargo has been approved for a determined period.

Name of the Candidate, Month Year

Statement of Authorship

By signing below I confirm that the work embodied in this thesis contains following published paper work of which **Ayesha Sahar** is the leading author. Chapter 4 contains parts of the paper "Tolerogenic Dendritic Cell reporting: has MITAP made a difference?" [Sahar et al., 2023]. It is worth mentioning that **Ayesha Sahar** had the active role in every stage of the study design, data collection and analysis, result analysis and manuscript preparation.

I included this written statement, endorsed by all authors, attesting to my contribution to the joint publications.

Name of the Candidate, Month Year	Collaborator 1, Month Year
Collaborator 2, Month Year	Collaborator 3, Month Year

Acknowledgements

I would like to say special thanks to my first supervisor Dr. Phillip Lord whose expertise and continuous support over the last three years have been invaluable. I am indebted to his encouragement in every thick and thin situation during my PhD studies. I am grateful to my other supervisors, Prof. Catharien Hilkens for giving me a diverse PhD experience with her expert knowledge in immunotherapies. I further extend my thanks to Dr. Jennifer Warrender who was my third supervisor.

This journey was supported by the "INsTRuCT" project funded by the European Union's Horizon 2020 research and innovation ITN program under the Marie Sklodowska-Curie grant agreement No. 860003. Due to this prestigious fellowship for PhD studies and my supervisors' support, I got many training opportunities for my career development.

During my PhD, I had an opportunity to do a secondment at SciBIte, Cambridge. I am grateful to them for giving me this excellent opportunity to work in a company. I am specially thankful to Michael Hughes, Lee Harland and James Malone.

Last but certainly not least, I am eternally thankful for the unwavering support of my husband throughout this challenging journey. Without him, I would not have been able to accomplish this.

Dedicated to,

All the incredible women of the world and especially to my dear daughter, Ayzel!

Her arrival during the final year of my PhD redefined my priorities and reshaped my perspectives, making the completion of my thesis not just an academic achievement but a personal milestone that I will cherish forever...

Contents

List of Figures

List of Tables

Abstract

In this thesis, we use data integration and analysis methods and examine the impact of data standardisation to enhance our understanding of tolerogenic dendritic cell (tolDC) therapies. Standardisation and structuring of the data are extremely valuable for it to be useful and accessible. Emerging biological fields face unique difficulties, including limited data availability, a lack of standardisation and challenges in knowledge management from different studies due to varied methodologies. These issues demand the development and application of specialised techniques and strategies tailored to their specific data handling and management needs.

This thesis focuses on one such emerging field, "tolerogenic Dendritic Cell Therapy", which has demonstrated significant potential. Like all biomedical experiments, developing these therapies involves several crucial steps that must be well-documented for comparison and replication purposes. Reporting frameworks, like Minimum Information Models can aid in standardising these descriptions; Minimum Information about Tolerogenic Antigen-Presenting cells (MITAP) was created in 2016 in this field for this purpose. We evaluate MITAP's impact on the field of tolDC therapies by analysing a selection of literature. We found that MITAP is utilised in a minority of relevant papers (14%), but where it is applied, there is slightly more metadata available. This suggests that while MITAP has had some success, further efforts are needed for standardised reporting to become widespread in the discipline.

In order to further aid the comparison, re-purposing and re-use of data about tolDC therapies, we built a method to identify and integrate the most significant information related to tolerogenic dendritic cell therapies into a knowledge graph structure. A key aspect of the knowledge graph is ensuring that the merged data is relevant to the field. We employ knowledge extraction techniques to identify and collect relevant information from

research articles, integrating this with publicly available datasets to enrich the knowledge base.

We successfully embedded this data into a comprehensive knowledge graph comprising 120k entities extracted from full-text articles and additional integration of 92k relationships from other relevant databases. The use of knowledge extraction techniques from research articles ensured the relevance of the integrated data to the field. It also allowed us to gain more insights from publications with unpublished experimental data, as shown in the example queries. This knowledge graph can act as a base for the generation of further hypotheses as well as a database for the storage and retrieval of relevant information about tolDC therapies.

Having built the knowledge graph our focus shifts to considering queries about the tolDC therapies that give us a better understanding of the degree of standardisation, about the underlying biology and the social environment in the field. We formulated diverse queries encompassing heterogeneity concerns. The results demonstrated the effectiveness of tolKG in promptly addressing these queries, a task that would either necessitate specialised expertise or significant manual scrutiny if pursued conventionally. Through the utilisation of tolKG, we streamline tasks such as comparison and analysis and even facilitate the generation of novel hypotheses.

In summary, we found that a knowledge graph is an effective way to integrate data. Moreover, the addition of data from the literature makes it more meaningful, especially for emerging fields where there is a lack of experimental data sharing. Text mining from literature enables the extraction of more relationships that are specific to a field. As a result, it can help to perform an effective analysis and comparison of the tolDC therapy field.

Together, this work helps establish the groundwork for applying data science methods in tolDC therapies making several kinds of comparisons possible which are not possible without it. The methodologies employed are specifically tailored to the data sources of tolDC therapies. Nonetheless, these strategies are not restricted to this particular domain; they primarily depend on the input data sources, which makes them usable in other areas of biology as well.

1

Introduction

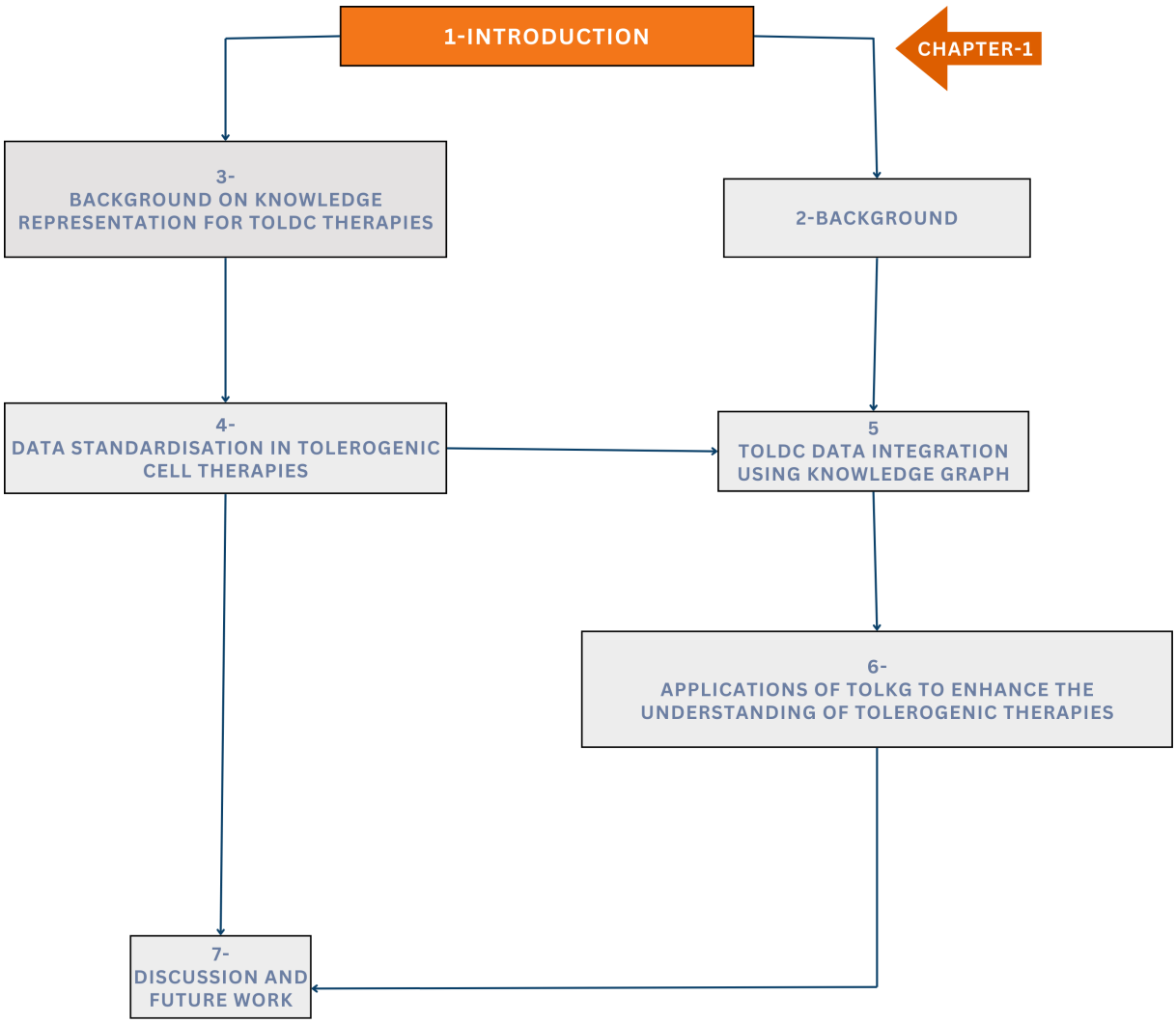


FIGURE 1.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

1.1 Overview of the thesis

"In God we trust; all others must bring data." - W. Edwards Deming

In this thesis, we are interested in advancing the data management and knowledge representation for a specialised area within immunology known as Tolerogenic Dendritic Cell (tolDC) therapies. Biological data management addresses the challenges of collecting, storing, organising and retrieving vast amounts of biological data. On the other hand, biological knowledge representation seeks to model and encode complex biological concepts, relationships and theories in a structured format. Moreover, advanced computational analysis and comparison can be performed by translating the available biological information into machine-readable formats. While the fields of biological data management and knowledge representation have distinct objectives and methodologies, they are united in their fundamental role: they underpin the efficient handling, interpretation and application of biological information. This integration is critical in facilitating advanced computational analyses and comparisons. To conclude, the synergy between biological data management and knowledge representation is pivotal not just for maintaining data integrity but also for enabling innovative discoveries in tolDC therapies. This thesis aims to contribute to this area by developing methods that not only streamline data processes but also enhance the utility of information, ultimately accelerating advancements in immunological treatments.

Here, we focus on data standardisation and data integration of the tolDC therapies. We explore the impact of a standardisation method known as MITAP (Minimum Information for Tolerogenic Antigen Presenting Cell) in the field and look at the reason for low uptake, suggesting ways to increase it. We are mainly interested in first finding the most suitable data to represent the field and then transforming it into knowledge. We define the data, information, knowledge and wisdom according to the DIKW paradigm, often represented as a pyramid, which stands for Data, Information, Knowledge and Wisdom [Gajzler, 2016].

The task of extracting knowledge from existing data sources is less time-consuming and less expensive compared to generating new data. This advantage is particularly notable in fields such as medicine, where conducting experiments can be costly due to many reasons including human tissue availability. However, due to the heterogeneity in the experiments, it is easy to be overwhelmed and accumulate data which is not relevant — a problem that

becomes more pronounced when the data is smaller and more sparse. This is the case in the field of tolDC and so, in this thesis, we explore the targeted integration of data into a comprehensive knowledge graph format, with a particular focus on a relatively nascent and smaller field, known as tolDC therapies.

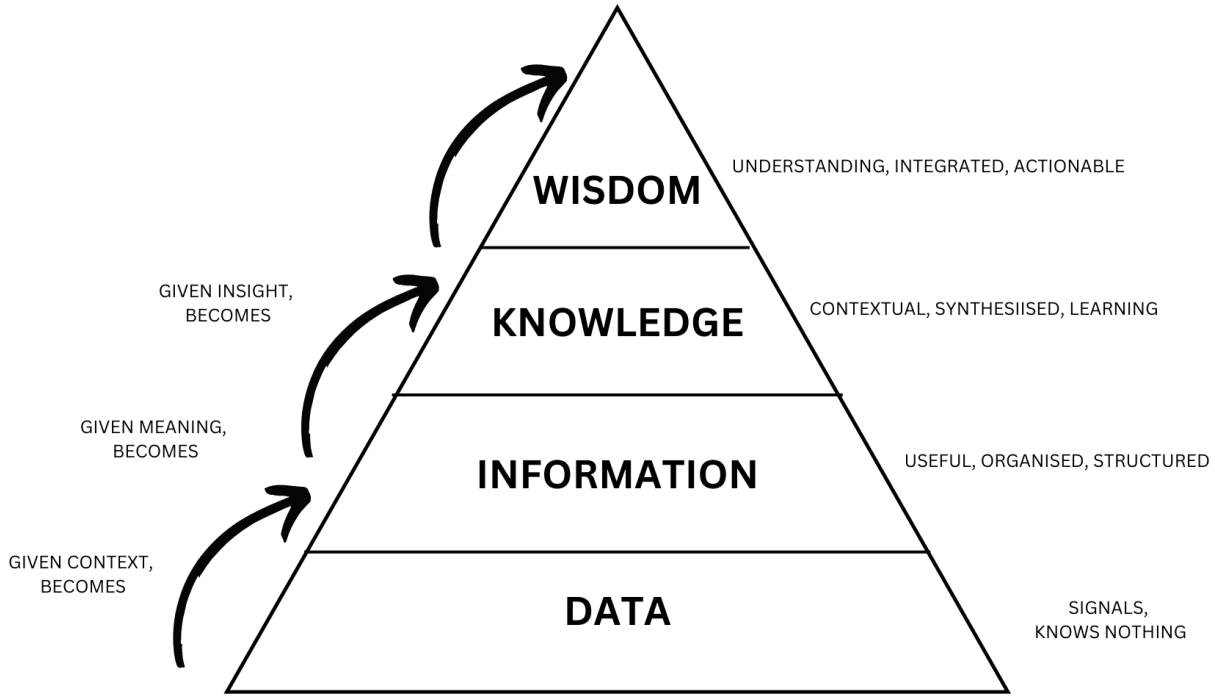


Figure 1.2: DIKW Pyramid: Data, Information, Knowledge, Wisdom [Gajzler, 2016]

1.2 Motivation

According to a recent UK population-based study, out of 22 million people studied from 2000 to 2019, roughly 4.4% were newly diagnosed with an autoimmune disease. Of those diagnosed, approximately 64% were women and 36% were men [Conrad et al., 2023]. According to a report published by Connect Immune Research Consortium in 2018 in the UK, a total of four million people are estimated to be living with at least one autoimmune condition. The financial burden is staggering, with just three of these conditions—type 1 diabetes, rheumatoid arthritis and multiple sclerosis, costing the UK more than £13 billion annually in direct and indirect expenses [Garcia, 2018]. With over 80 known autoimmune conditions, the issue is far-reaching. Alarmingly, the incidence of many autoimmune conditions is on the rise, some increasing by as much as 9% each year [Lerner et al., 2015]. In this context, tolDC therapies emerge as a promising treatment. These therapies aim to modulate the immune system in a way that reduces its attack on the tissues of the body, offering a targeted approach that could alleviate symptoms and potentially reduce the overall social and economic burden of autoimmune diseases.

However, the field of tolDC therapies is relatively new and suffers from low data reporting and standardisation issues, as discussed in Chapter 3. We can enhance the progress of the field by employing integrative bio-informatics in this area. Historically, the biological scientific community has suffered from a *data reproducibility crisis*; this refers to the growing concern about the lack of consistency in experimental results when independent researchers attempt to reproduce them [Begley and Ellis, 2012]. Yet, in recent times, there has been a lot of interest in data reuse and the adoption of the FAIR data principles [Munafò et al., 2017]. As a result, the introduction of computational and integrative data methodologies has profoundly transformed numerous scientific disciplines, especially those with a more extensive and established foundation. However, it is equally, if not more, crucial for smaller and emerging fields to leverage these advanced methodologies. The benefits are manifold, including accelerated discoveries and advancements in the respective fields.

In this thesis, we aim to transition from the challenges of inadequate data reporting and standardisation towards creating an extensive knowledge repository, specifically centered on tolDC therapies. By transforming limited raw data into meaningful insights, we can accelerate the pace of new discoveries.

1.3 Research aims and questions

As supported by the literature review (see Chapter 2), basic and translational research in tolDC therapies could be significantly advanced through better reporting of experiments, which enables comparison, standardisation and reproducibility. This project aims to combine data about tolDC therapies in a comprehensive manner, thereby facilitating reproducibility, comparative analysis and meta-analysis within the field. The main aim of the project was:

"Use data science approaches to characterise the tolDC therapies in a standardised manner"

1.3.1 Research questions

The main research questions of this thesis are:

1. **RQ1** How can the standardisation be improved and promoted in the tolDC therapies field?
2. **RQ2** How can the Integration of all tolerogenic therapy data in one place help the standardisation, harmonisation and visualisation of data thus empowering translational research in the field?
3. **RQ3** What level of heterogeneity exists in tolDC experiments and the social aspects of the tolDC field? To what extent is this variation beneficial?

1.4 Research contributions

The main contributions of this thesis are investigating and promoting the usage of standardisation models in the tolDC therapies field and a comprehensive targeted knowledge graph built specifically for tolDC therapies. The following are the parts of the developed approaches that answer the research questions and form the contributions of this thesis.

- **Promoting the standardisation for the reporting of tolerogenic dendritic cells data reporting**

Standardisation protocols exist within the field, yet their adoption across all research groups is not universal. As tolDC therapies approach the final stages of clinical trials, the implementation of these standardisation protocols is becoming increasingly crucial. A minimum information model, in this context, is a set of guidelines that stimulate the minimum amount of information required to ensure that the results of an experiment can be easily understood, reproduced and compared with similar experiments. An investigative study was conducted in Chapter 4 to ascertain the percentage of researchers employing the existing minimum information model. This study yielded insightful findings concerning the use of minimum information models, shedding light on both specific trends within the field of tolDC therapies and broader usage patterns across different research groups, thus fulfilling **RQ1**.

- **Construction of tolKG**

  As part of the literature work, we investigated the availability of data in the tolDC field and discovered that the limited data available is difficult to access due to the absence of a dedicated tolDC database. Existing immunology databases contain little to no data on tolerogenic dendritic cells. Given the absence of a dedicated database providing data specific to tolDC therapies, the initial objective is to integrate data for tolDC therapies. This integration incorporates pertinent background resources, including pathway data, cell development data and pharmacological datasets. The expected sources of data are shown in Figure 1.3. The successful execution of this integration paves the way for reproducibility and comparison within the field, thereby addressing **RQ2** and **RQ3**.

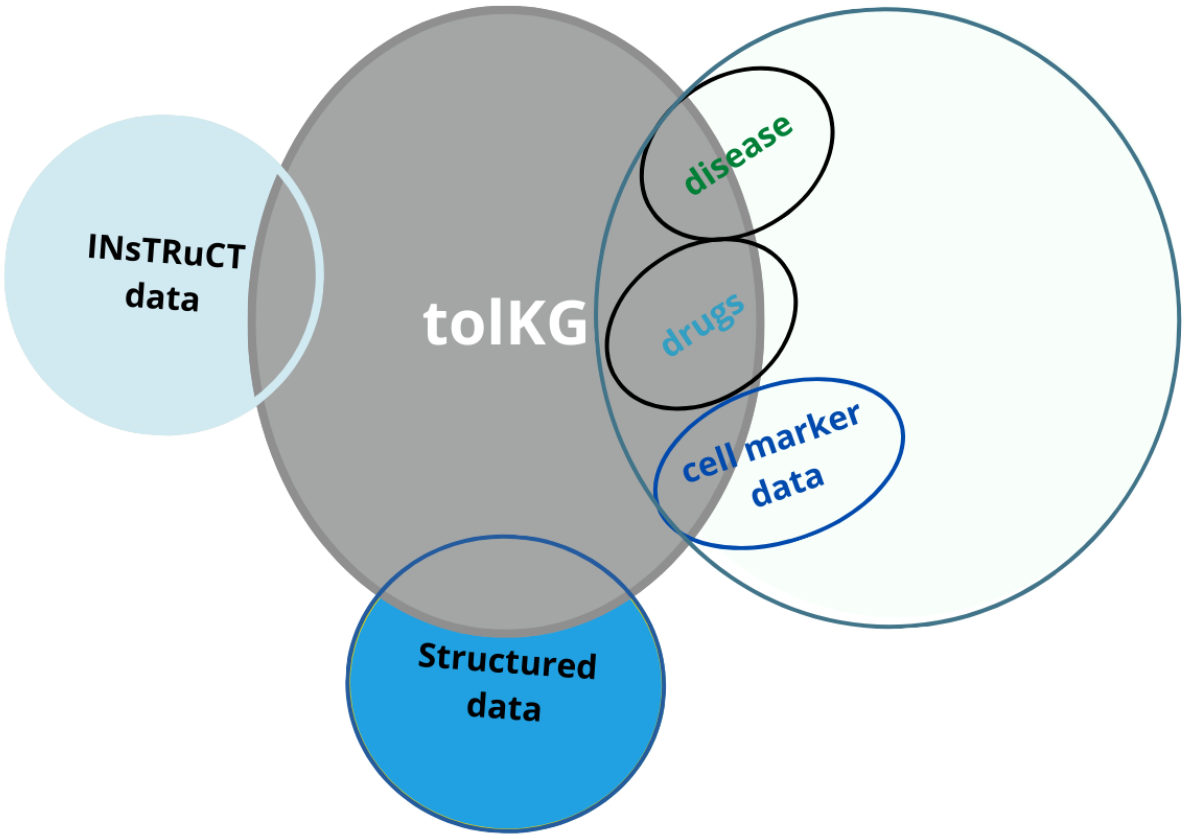


Figure 1.3: Data sources and types
Diagram to show the expected sources and types of data for integration into tolKG

The development of this targeted database consists of a series of procedures dedicated to the systematic compilation, structuring and preservation of data pertinent to a distinct field as shown in Chapter 5. In niche fields, such as tolDC therapies, there exists a conspicuous dearth of readily available online data. Under these circumstances, it becomes essential not only to utilise extant datasets but also to ensure the relevance of the incorporated data. To this end, we have established a data integration pipeline that optimally leverages both available data sources and information gleaned from scholarly literature. Importantly, this pipeline bears the potential for application across other nascent or specialised research fields as well.

- **Understanding the heterogeneity in tolDC field**

  After developing tolKG, we use it to gain a deeper understanding of the tolDC field. While various tolDC protocols are known to exist, the full extent of their diversity remains unclear. In Chapter 6, we form different queries to investigate the breadth of variation among these protocols. We start by exploring the use trend for inducing agents and growth media in the tolDC generation. Furthermore, we investigate the preferences of researchers for maturing tolDC cells. Finally, we seek to understand if the tolDC produced with different agents differ significantly in their outcomes and if one type of tolDC is more suitable for a certain category of diseases than others. In the end, we also perform a social analysis of the field in terms of gender bias and trends. All these questions vary in their complexity and are not possible to be answered without tolKG. They help us understand the heterogeneity in the field in a better way, thus fulfilling **RQ3**

1.5 Organisation of the thesis

The structure of this thesis is designed to follow the natural progression of the research process, moving from foundational knowledge to application and evaluation. **Chapters 1 to 3** provide the context and background, establishing the importance of data standardisation and integration in tolDC research while reviewing existing methodologies. **Chapters 4 and 5** present the core methodological contributions, focusing on the analysis of MITAP's impact and the development of tolKG as a structured approach to data integration. **Chapter 6** shifts towards evaluation and application, demonstrating how

tolKG can be utilised to answer real-world research questions, highlighting its practical benefits. Finally, **Chapter 7** serves as a synthesis of findings, reflecting on the challenges encountered, discussing the broader implications of the work and outlining potential directions for future research.

- **Chapter 2:** This chapter presents brief backgrounds to the different topics that are discussed later in the thesis. It starts by introducing the motivational background for the emergence of knowledge representation. It then introduces some of the latest data representation frameworks in the biomedical domain. It finally provides a brief background on the semantic data representation.

- **Chapter 3:** This chapter presents literature-based approaches to represent data and knowledge in subfields such as tolDC therapies. In this chapter, we discuss the need for data representation for tolDC therapies and identify the challenges related to data in the tolDC therapies. This chapter then discusses text mining and standardisation approaches for data as they are the initial steps in data representation. It then discusses different approaches to forming a database that can be used for knowledge representation. These approaches include our pioneered graph-based data integration.

- **Chapter 4:** This chapter presents standardisation approaches in the field of tolDC therapies. After introducing the minimum information model, it briefly explains a minimum information model that was developed for tolDC therapies. Then, it presents an extensive analysis of the usage of that minimum information model. Finally, this chapter also highlights some important facts about the usage of minimum information models; in the end, it discusses the steps that can be taken to promote the usage of standardisation approaches, specifically the minimum information models.

  Publication Status: The results of this chapter were published in PeerJ journal, examining the impact of the Minimum Information Model on tolerogenic therapies research and its broader adoption across other biomedical fields [Sahar et al., 2023].

- **Chapter 5:** This chapter presents a graph-based approach for the integration of data related to tolDC therapies. After explaining the unique challenges faced by

subfields for data integration. it explains a comprehensive framework for data integration for tolDC therapies. This framework includes data mining from text as well as other relevant databases. It also discusses the workflows and pipelines for information retrieval from text. In the end, the chapter shows how the data can be integrated into a knowledge graph. Then, the chapter concludes with a discussion on different aspects of the framework.

Publication Status: Sections of this chapter have been presented in an international conference on knowledge graphs. Parts of this chapter have been incorporated into a paper which was submitted to Biodata mining journal. We recieved the feedback on the paper and plan to re-submit it in a more suitable journal.

- **Chapter 6:** This chapter presents the evaluation and real-world use cases of tolKG. Following an introduction to the types of queries that offer valuable insights for researchers via tolKG, we categorized them into distinct groups. The initial category involves utilizing tolKG to streamline the sorting of research studies, while the subsequent one pertains to leveraging the capabilities of tolKG to tackle heterogeneity within the tolDC therapies domain. Ultimately, the chapter wraps up with a comprehensive discussion encompassing various facets of the framework.

  Publication Status: Sections of this chapter have been incorporated in the evaluation part of the tolKG paper which we plan on resubmitting to a suitable journal.

- **Chapter 7:** This chapter reviews the outcomes of this work in the broader context of applicability in the field and describes possible future work.

2
Background

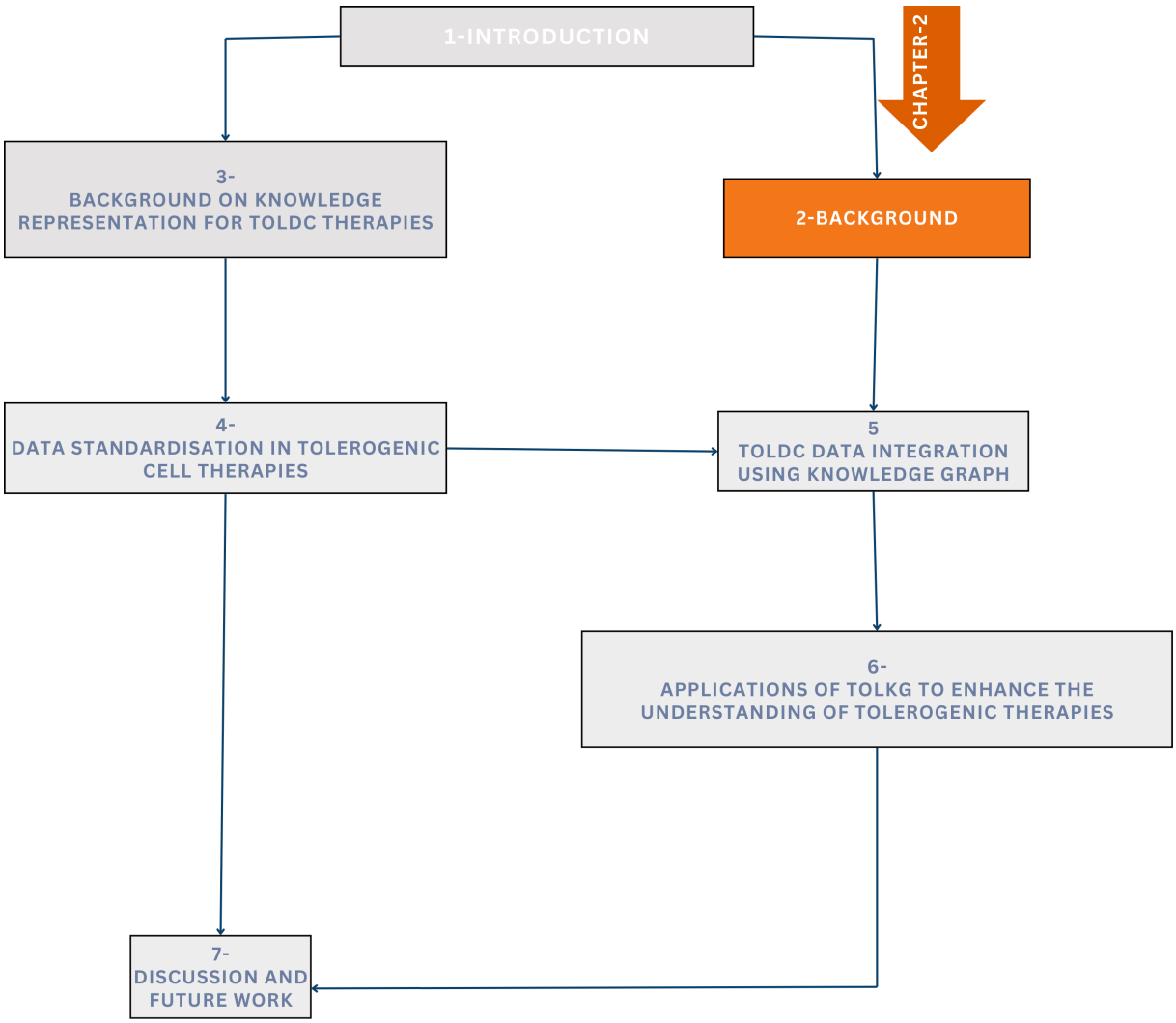


FIGURE 2.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

2.1 Greater context

The progress of biology has been marked by advancements in data representation and storage throughout history. For instance, Carl Linnaeus introduced a standardised binomial naming system for categorising species [Linnaeus, 1758]. This system provided a standardised way of naming and organising species, laying the foundation for the study of biodiversity and facilitating the comparison of organisms across different regions and periods. The naming scheme has had an impact far beyond the taxonomy that he produced, which has been largely discarded.

The importance of data visualization was further highlighted by Florence Nightingale in the 1850s through her innovative use of the Rose Diagram to depict mortality causes during the Crimean War. This early example of visual data representation emphasized its utility in scientific communication and decision-making. Moving into the modern era, the advent of sequencing technologies brought about a paradigm shift, requiring new approaches to data management. For example, the genome sequencing of various organisms created a vast influx of genetic data, leading to the creation of specialised databases and bioinformatics tools that have transformed how genetic information is understood and analysed.

These advancements ultimately led to groundbreaking endeavours such as the Human Genome Project which achieved a complete mapping of human genes and demonstrated the intertwined evolution of biological sciences and data representation techniques [Olson, 1993]. Together, these milestones illustrate a continuous trajectory in biology, where the capability to manage and represent data effectively has been central to advancing our understanding and facilitating comparisons of biological complexity throughout history.

In summary, we learn from these examples that the advance of biology has been characterised by its evolving methods of data representation and storage. From the binomial naming scheme to genetic ratios, species taxonomy and modern DNA and genome sequence representation, these advancements have played a critical role in unravelling the complexities of life. As technology continues to progress, data representation in biology still remains essential for making new discoveries, understanding biological systems and addressing the challenges of our rapidly evolving world.

2.2 Characteristics of databases

In this section, we consider the nature of databases, by building an informal set of characteristics, and comparing between biological and non-biological databases. As far as we are aware there is no standard classification for these characteristiscs.

We have just argued that data representation, standards and handling play a crucial role in the advancement of biology. Next, we consider how data is stored in biological databases. Here, we identify some key characteristics of the different types of databases.

- **Data size:** Data size refers to the amount of space that a particular set of data occupies.
- **Openness:** Generally speaking, openness refers to the ease of public accessibility, freedom from proprietary restrictions and the transparency of the data and its collection processes.
- **Data Complexity:** Biological data can be highly complex, involving not only raw sequences but also associated metadata, annotations and references. For example, a gene sequence in a biological database may include information about its function, expression and related diseases. Other databases may have simpler data structures.
- **Volume of traffic:** It refers to the number of transactions or interactions that take place within these platforms. These transactions can include things like data queries, data additions or updates.
- **Read/write:** Read operation refers to retrieving or accessing data from a database. Write operation refers to adding new records or data to the database. Here, we are interested in how many people update the database and how many people read it.
- **Data federation:** Data federation is a data integration approach that allows organisations to access and query data from multiple, disparate sources as if they were a single, unified source. This approach is often used in situations where an organisation has data spread across various databases, systems or data warehouses and needs to aggregate and analyse that data without physically moving or replicating it.

- **Users**: Biological databases are primarily used for biological research, including genomics, proteomics, structural biology and bioinformatics. Other databases serve diverse purposes, such as managing business operations, financial transactions or geographic mapping.

We have applied these characteristics to multiple databases. However, a rigorous analysis is not always possible due to data limitations. Nevertheless, this approach offers a broad understanding of where biological databases are positioned.

From the heatmap, shown in Figure 2.2, we can quickly compare the databases across multiple attributes. Biological databases like the 10,000 genome, Uniprot, Immport and Gene Bank present a different profile. They might not handle as many transactions as medical databases, but they stand out for their high "Openness" scores, suggesting a greater degree of public access or open-source data availability. Their user base is more specialised and their data interaction, in terms of reading and writing, is moderate. Interestingly, while their complexity is generally lower than that of medical databases, their data federation capabilities vary, with Gene Bank leading the pack.

On the other hand, medical databases, such as WHO, UK biobank, CPRD and NHS data, are characterized by their high scores in the "Transaction" and "Users" categories, signifying their capability to manage a vast number of transactions and cater to a large user base. They also exhibit significant data interaction, as evidenced by their elevated "Read/write" scores. While their openness varies, with WHO and UK biobank being more open compared to CPRD and NHS data, they generally possess intricate structures and are likely integrated with other data sources, as indicated by their moderate to high scores in "Complexity" and "Data Federation".

Lastly, the category of other databases, which includes giants like Google, Facebook, Bank data and NASA, underscores their vast scale. They dominate in terms of size, transaction rates and user base. However, their openness is more restricted, especially in the cases of Bank data and NASA. Their data interaction patterns, as reflected by the "Read/write" scores, are diverse, with some like Google and Facebook being more active than others.

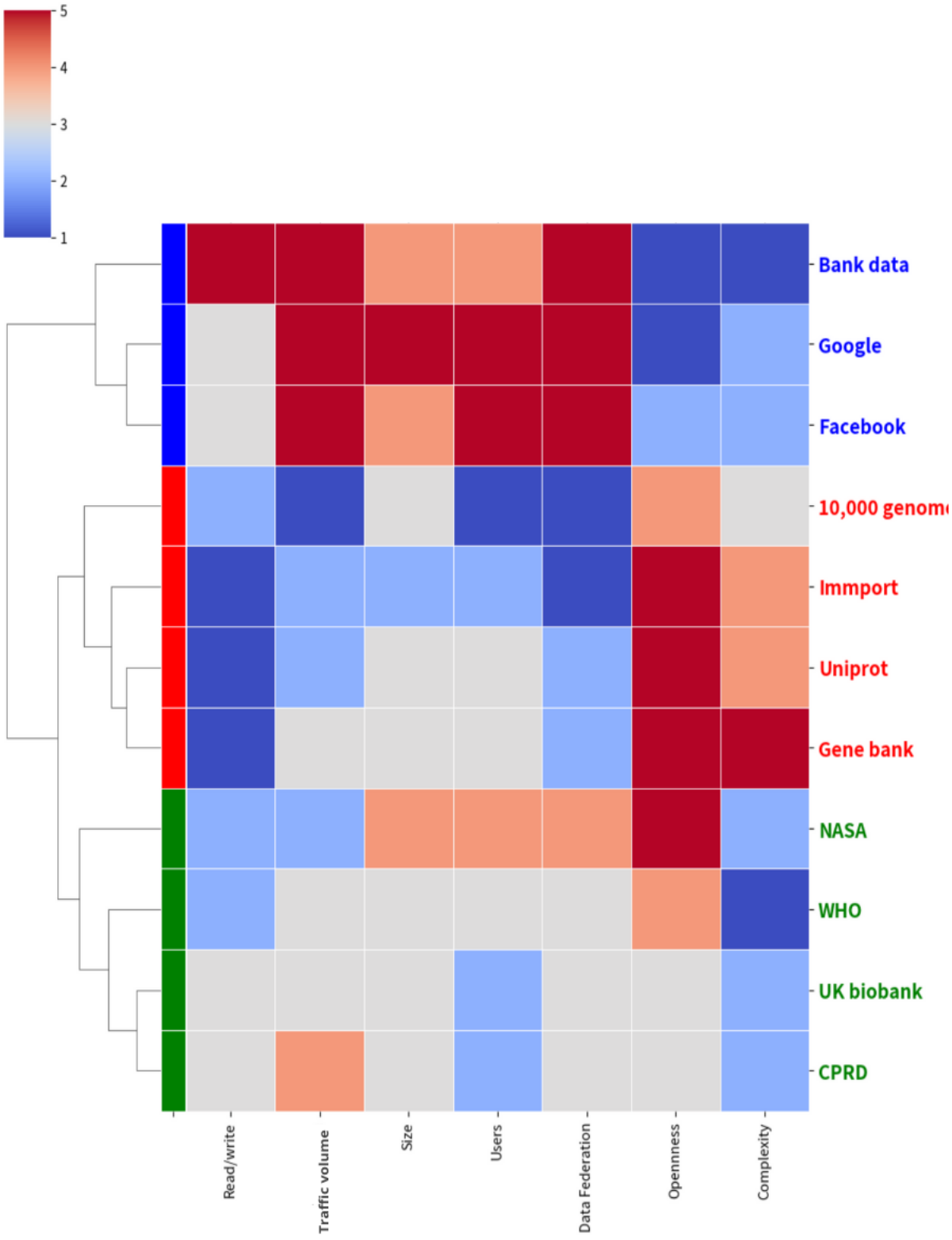


Figure 2.2: Comparison of different types of databases
The values range from 1, indicating the lowest, to 5, indicating the highest. The data presented in this figure are approximate estimates derived from Google searches and reflect the general context.

2.3 Comparison between different biological databases

As noted in the previous section (Section 2.2), biological databases are different from the other categories as they have a specialised audience and need to be open access as well as standardised. Here, we compare different biological databases to see how each is customised to meet specific research demands and challenges. We have selected the biological databases which are recent and also relevant to the work done in this thesis.

In the above section, we noted that the biological databases deal with specialised researchers but here we look at whether the database is developed for a specific subfield or serves more general areas in biology. We will also assess whether they include heterogeneous data and utilise text mining, both vital for effective knowledge representation in biology. Additionally, although we established that biological databases should be openly accessible, we will now evaluate whether all databases in this category meet this criterion by checking their online availability, whether they feature standardised entities, and if they are graph-based, indicating whether data integration is performed using graph databases or relational databases. We will discuss more on the benefits of using graph-based data integration later in this chapter.

Table 2.1: Comparison of the recent relevant work. ✗means no ✓means yes and ❍partially.

Database	**Specific sub-field**	**Hetero-genous**	**Text mining**	**Online**	**Standard-isation**	**Graph based**
LinkedImm	✗	✓	✗	✓	❍	✓
Immport	✗	✓	✗	✓	❍	✗
Innate-DB	✗	✓	✗	✓	✓	✓
Reactome graph DB	✗	✗	✗	✓	❍	✓
DISEASES	✗	✗	✓	✓	✓	✗
Gene Expression Omnibus	✗	✓	✓	✓	✓	✗
Drug Repos-itioning	✗	✓	✗	✗	✓	✓
Drug-CoV	✓	✓	✓	✗	✓	✓

Biological databases often utilise standardised formats and standard ontologies (GO, Pathway Ontology) to ensure data interoperability and usability across platforms such as standardised ontologies are used in LinkedImm and Immport. The level of standardisation in databases like LinkedImm, InnateDB and GEO (Gene Expression Omnibus) varies based on their design, data types and specific goals. Here, we say that a database has followed standardisation if it is highly standardised, like using Highly standardized submission formats (SOFT, MINiML, MIAME, etc.)

These variations in standardisation and structure are particularly relevant when we consider how databases support specific fields of research, such as immunology, drug repurposing or Covid-19. As many biological databases are designed for specialised research areas, it is important to understand how they serve the needs of specific disciplines. The following sections illustrate how different databases are tailored to address the complexities of types of data integration and research demands.

Immunology: As shown in Table 2.1 Immport, Linked-Imm and Innate DB are the immunology databases. ImmPort (Immunology Database and Analysis Portal), shares data from clinical trials and research studies to accelerate the discovery of new diagnostic tools and therapies [Bhattacharya et al., 2018]. It includes data from diverse research studies, ranging from basic immunology to clinical trials with the aim to make the data publicly available to researchers. However, Linked-Imm aims to provide an integrated view of heterogeneous datasets such as genes, pathways and transcriptional profiling from other databases, including Immport, which is why it has been developed using a graphical database model [Bukhari et al., 2019]. Similarly, InnateDB is focused on genes, proteins, experimentally-verified interactions and signalling pathways involved in the innate immune response and is also developed using a graph-based framework [Breuer et al., 2013]. This shows that the purpose of the database plays an important role in deciding how it should be developed.

Drug repurposing: In recent years, research into finding new uses for existing drugs has hugely benefited from the growth of large biomedical databases [Banerjee et al., 2020]. Many computer-based methods have been created to thoroughly analyse different types of biomedical data. One of the crucial steps for drug repurposing is the efficient integration of several kinds of datasets to explore the relationships. Several researchers are focused

on the integration of data for drug repurposing such as development of a semantically-rich drug discovery network for drug repurposing [Mullen et al., 2016]. Utilising this network, the researchers highlight central nervous system disorders and mine the network to infer and rank potential drug-disease relationships not previously captured. The network is created using graph graph-based database to capture the semantic associations.

Covid-19: One of the recent examples where data integration played an important role was during the Covid-19. Many researchers worked on finding the cure and intervention for the virus using already available resources. For instance, Drug-CoV is a knowledge graph specifically designed for drug repurposing in the context of COVID-19. It integrates structured data from several public databases (such as DrugBank, PubChem and MedlinePlus), focusing on identifying existing drugs that could potentially be repurposed to treat COVID-19. The graph includes entities like drugs, genes, proteins, diseases and side effects, all connected by multi-relational links. Drug-CoV aims to accelerate drug discovery by identifying meaningful drug-disease relationships based on biological interactions and prior clinical trials [Li et al., 2023]

COVID-19 KG integrates and unifies various COVID-19 datasets into a single, semantically consistent and machine-readable knowledge graph [Domingo-Fernández et al., 2021]. It was designed to support research and applications related to COVID-19, including drug repurposing. Similarly, Covid-19 Disease Map was a collaborative effort focused on curating and collating molecular interaction mechanisms of the virus into a computable, comprehensive diagram, which was made available in various data formats and semantic standards, enabling computational analysis and modeling [Ostaszewski et al., 2020]. Moreover, research demonstrated the application of semantic data integration for COVID-19 scientific publications [Chen et al., 2021]. The study demonstrated that this approach was useful for quickly grasping the current state of medical knowledge and discovering relationships between emerging medical concepts. The authors also mention an extraordinary volume of relevant medical literature produced during the global pandemic and the potential for further development of this approach in analysing large collections of literature.

2.4 Natural Language Processing (NLP)

Natural Language Processing (NLP) is a subfield of artificial intelligence and computational linguistics that focuses on the interaction between computers and human (natural) languages. NLP aims to enable machines to read, understand and derive meaning from human language in a way that is both valuable and useful. It covers a wide range of tasks, including text analysis, language translation, sentiment analysis and the extraction of entities and relationships from large text corpora. These tasks are vital for making unstructured text data more accessible and interpretable for various applications, such as search engines, virtual assistants, and automated customer support systems.

Named Entity Recognition (NER) is a core task in NLP that involves identifying and categorising key pieces of information, known as "entities", from unstructured text. These entities could be names of medical terms or other specific data points relevant to a particular domain. The goal of entity extraction is to transform unstructured text into structured data by identifying and labelling the relevant components within it, making the data more usable for tasks like search, analytics, and data integration [Nadeau and Sekine, 2007].

For example, in the medical domain, entity extraction is widely used to pull out clinical concepts from patient records or biomedical literature. Consider a sentence from an electronic health record (EHR) such as:

"The patient was diagnosed with type 2 diabetes and prescribed metformin."

In this case, NLP techniques can be applied to extract entities such as:

- Disease: type 2 diabetes
- Medication: metformin
- Patient action: diagnosed, prescribed

Once extracted, these entities can be categorized into predefined classes such as medical conditions, treatments, and patient behaviors. This process is crucial for converting unstructured clinical notes into structured formats that can be easily queried and analyzed. Studies have shown that entity extraction in medical records improves clinical decision-making by organising information into structured and searchable formats [Savova et al., 2010].

Having examined the characteristics and comparison of various databases, the next focus is on the relevant techniques employed in building these databases, including text mining and data integration.

2.5 Semantic data representation

Semantic data representation is defined as a meaning-oriented approach that utilizes subjective views of entities and relationships to describe semantic information, establishing a graph-based model to specify taxonomy and compound construction of concepts [Li, 2016]. One approach defines semantic data representation in biomedical research as the use of Semantic Web Technologies (SWTs) such as the Resource Description Framework (RDF) and the Web Ontology Language (OWL) [Bona et al., 2019]. The World Wide Web Consortium (W3C) plays a crucial role in formulating these standards that enhance semantic data representation, such as RDF, OWL and SPARQL, which is a query language for RDF. These standards facilitate data modelling, representation and querying, enabling semantic interoperability and the extraction of multimedia information [Merabti et al., 2012].

The field of Artificial Intelligence (AI) in the 1980s first demonstrated the utility of ontologies in knowledge representation, leading to the development of the Semantic Web by Tim Berners-Lee and others at the turn of the century [Berners-Lee et al., 2001]. With the establishment of the W3C standards for ontology languages (like RDF, RDFS, OWL), the use of ontologies in computer science has expanded, especially in data management [McGuinness, 2007]. Ontologies have been used widely in the biomedical domain to provide a structured framework to represent knowledge in a formal, understandable and machine-readable way [Smith et al., 2007]. Biomedical ontologies capture concepts (such as diseases, genes, symptoms etc.) and the relationships between these concepts (such as “is a symptom of”, “is caused by”, etc.). The use of ontologies in the biomedical domain supports data integration, annotation, interoperability and precise encoding in electronic health records (EHRs), facilitating a systematic approach to managing complex biomedical information [Vechina et al., 2013]. For instance, one study integrated the Disease Ontology (DO), Human Phenotype Ontology (HPO) and Radiology Gamuts Ontology (RGO) to explore causal relationships between high-level DO and HPO concepts, offering new insights into the connections between different classes of diseases and

symptoms [Kahn Jr, 2022]. This integration exemplifies how ontologies can enrich the biomedical research landscape by providing a comprehensive view of possible disease pathways and diagnostic criteria.

Although Knowledge Graphs are conceptually similar to ontologies, they are more recent in their development and usage. They were popularized by Google, who introduced their Google Knowledge Graph in 2012 to enhance their search engine performance with semantic-search information gathered from various sources [Singhal et al., 2012]. Recently, knowledge graphs have become popular among researchers in the biomedical domain. A Knowledge Graph in semantic data representation is a structured framework that organizes information into entities (nodes) and their interrelationships (edges) [Ehrlinger and Wöß, 2016]. This graph-based approach integrates diverse data sources and enables complex queries based on the rich, interconnected structure of the data. In essence, a Knowledge Graph uses the ontologies as a sort of "blueprint" to organise and structure its data. The ontology provides the rules and definitions and the Knowledge Graph applies those to specific instances of data.

In summary, while an ontology provides a theoretical and structured framework for understanding domain-specific knowledge, a Knowledge Graph uses such frameworks to manage and utilize actual data. An RDF graph, on the other hand, is a method of implementing these concepts, particularly in the context of the Semantic Web, using a standardised format to describe and interlink data.

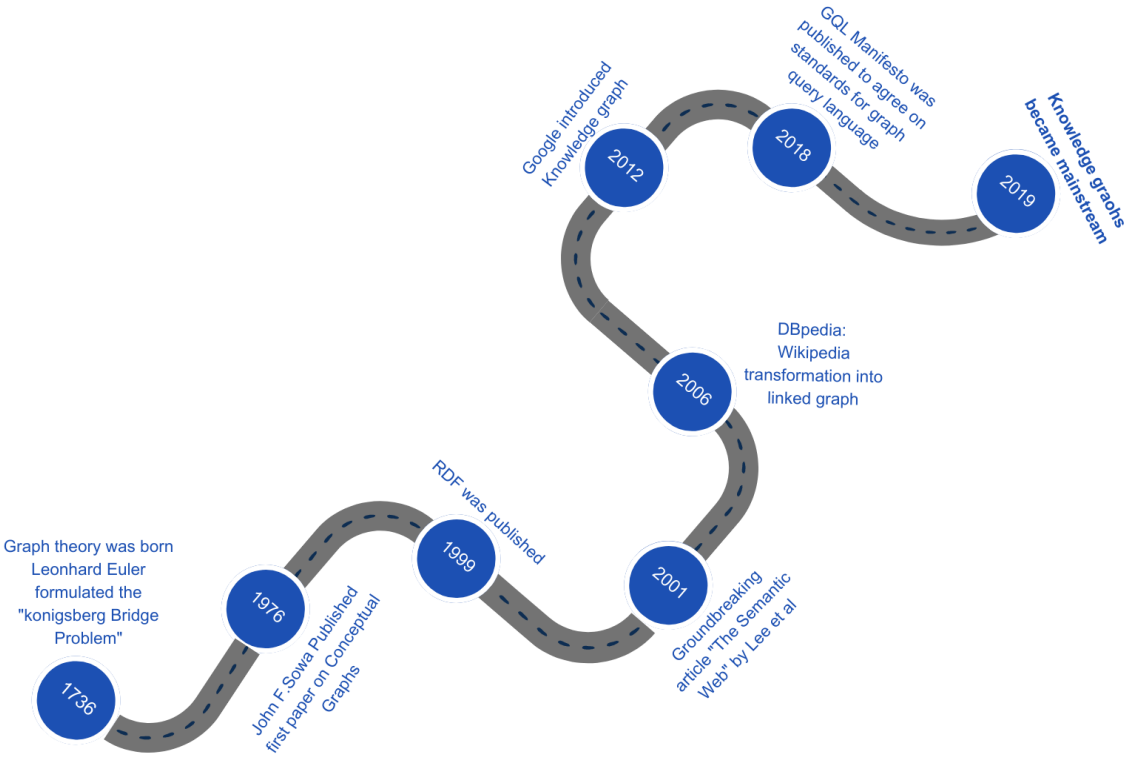


Figure 2.3: History of graph for representing knowledge

2.5.1 Neo4j

Neo4j is primarily known as a graph database that uses its own data storage model and query language called Cypher. It is designed to store and manage data as nodes, relationships and properties, which forms a graph structure that is highly efficient for representing complex networks and performing related queries.

The core strength of Neo4j lies in its ability to perform deep, complex queries at high speed. This capability stems from its native graph storage and processing engine, which is optimized for traversing relationships [Robinson et al., 2015]. As a result, Neo4j is widely used in various domains such as social networking, recommendation engines, fraud detection, network and IT operations and more, where the relationships between data points are as critical as the data itself [Park et al., 2014].

An RDF graph, in contrast, is a standard model for data interchange on the web, structured in triples (subject, predicate, object). RDF is particularly used in Semantic Web applications to link data with a variety of structures across different systems.

While Neo4j primarily supports its native graph data model, it can be used with RDF data through additional tools and plugins that translate RDF into a form that can be stored and queried within Neo4j. However, out of the box, Neo4j does not natively handle RDF in the same way that dedicated RDF stores (like Apache Jena or Virtuoso) do.

Cypher Query Language: Cypher is the query language for neo4j. It allows users to efficiently query and update graph data by expressing what they want to select, insert, update, or delete in the graph database without describing how to do it. Cypher syntax is designed to be readable and intuitive, making it accessible to developers and database administrators alike. Its pattern-matching capabilities enable users to visually describe patterns in their data and perform sophisticated queries over complex relationships.

Cypher queries are based on specifying patterns of nodes and relationships to search for in the graph. These patterns can be augmented with conditions on properties, allowing for precise and flexible querying capabilities. Cypher also supports aggregations, sorting and limiting results, providing a powerful toolset for analyzing connected data.

2.6 Discussion

Various formalisms, such as ontologies, semantic networks, frames and rules, are used to represent knowledge. It involves creating a structure to store information, defining relationships between entities and developing rules and logic that guide reasoning and decision-making. Data integration provides the raw material (data) that, when processed and structured properly, becomes knowledge. Knowledge representation takes this data and structures it into a format that can be used by AI systems for reasoning, learning and decision-making.

2.6.1 Lightweight data intgeration

Lightweight data integration refers to a streamlined approach to combining data from different sources with minimal complexity, lower costs and reduced resource requirements. It contrasts with traditional, heavyweight integration methods that might involve extensive infrastructure, complex transformations and significant resource investments.

Considering the challenges associated with integrating data from diverse sources, lightweight data integration is required to ensure data quality and accessibility. It offers a flexible, cost-efficient and scalable solution to handle the complex challenges of integrating heterogeneous data sources. By leveraging semantic technologies, it ensures that the data is not just integrated but also understood in its context, paving the way for more intelligent applications and insights.

2.6.2 Graph Based Data Integration

Using graph databases offers significant advantages for representing complex systems; a practice that finds wide application in computer science and network analysis (both social and technological) and is particularly crucial for bioinformatics research. These databases leverage semantic graphs where both the connections (edges) and the entities (vertices or nodes) are categorized with specific types from a predefined set and are further enriched with detailed attributes. This structured approach not only enhances the richness of the data representation but also facilitates its organization and accessibility. The defining characteristics of graph databases, such as their ability to efficiently store and interlink data from various sources, make them an excellent choice for data integration tasks. By

enabling seamless connections between disparate data points, graph databases greatly enhance the analysis and interpretation of complex datasets, thereby offering a powerful tool for researchers and practitioners across a range of disciplines.

However, despite these advantages, implementing and maintaining knowledge graphs presents several challenges. One of the primary difficulties is data heterogeneity as integrating information from different sources often requires extensive preprocessing, entity resolution and standardisation efforts. Additionally, KGs require specialised query languages such as Cypher as discussed in section 2.5.1. Another major challenge is scalability, as large-scale knowledge graphs with millions of entities and relationships can demand high computational resources and complex indexing strategies to ensure efficient querying. Furthermore, data quality and curation remain crucial concerns—incorrect or outdated information can propagate errors throughout the graph, potentially leading to misleading analyses.

2.7 Sumamary

The history of biology has been shaped by evolving methods of data representation and storage, from the Linnaean taxonomy to modern DNA and genome sequencing. Key landmarks include the standardised naming system by Carl Linnaeus, Gregor Mendel's foundational work on inheritance and the revolutionary discovery of DNA structure. Advancements in data standards and bioinformatics tools have enabled more comprehensive insights into genetic makeup and facilitated data sharing and analysis.

In addition to biology, computational and integrative data approaches have revolutionised various scientific fields, with an emphasis on the transformation of raw data into actionable knowledge. This transformation is imperative in areas like tolDC therapies, where smart and semantic-based approaches can extract meaningful insights from complex datasets, fostering the development of novel therapeutic strategies.

The state-of-the-art in data representation in the biomedical domain has seen advancements in heterogeneous data integration, with a focus on graph databases. Various databases serve different purposes, such as systems vaccinology studies, pathways analysis, mammalian innate immunity analysis and drug repurposing. The application of graph models simplifies data modelling and facilitates the exploration of relationships between biomedical entities.

Examples of data integration can be found in immunology, drug repurposing and COVID-19 research. Graph-based models and semantic associations have enabled new discoveries and interventions in these areas. The integration and unification of COVID-19 datasets into knowledge graphs and computable diagrams have particularly supported research and applications related to the pandemic.

Semantic data representation has brought a new dimension to structuring and interlinking data, making it meaningful to computers. The development of standards like RDF, OWL and SPARQL by the W3C has played a vital role. Biomedical ontologies, knowledge graphs and related technologies have become popular for organizing information into entities and relationships. This has implications for data management, natural language processing and AI, with Knowledge Graphs being popularized by companies like Google to enhance semantic search.

In conclusion, the chapter highlights the significant progress made in the understanding and application of data representation in biology and biomedical research. The integration of heterogeneous data sources and semantic information structuring has not only advanced our understanding of complex biological systems but also opened up new avenues for scientific exploration and innovation. Our goal is to use this experience from these advanced data representation techniques to bear in the emerging field of tolDC therapies. However, the field of tolDC therapies is nascent and faces challenges for data representation which are discussed in the next chapter.

3

Background to knowledge representation in tolerogenic dendritic cell therapies

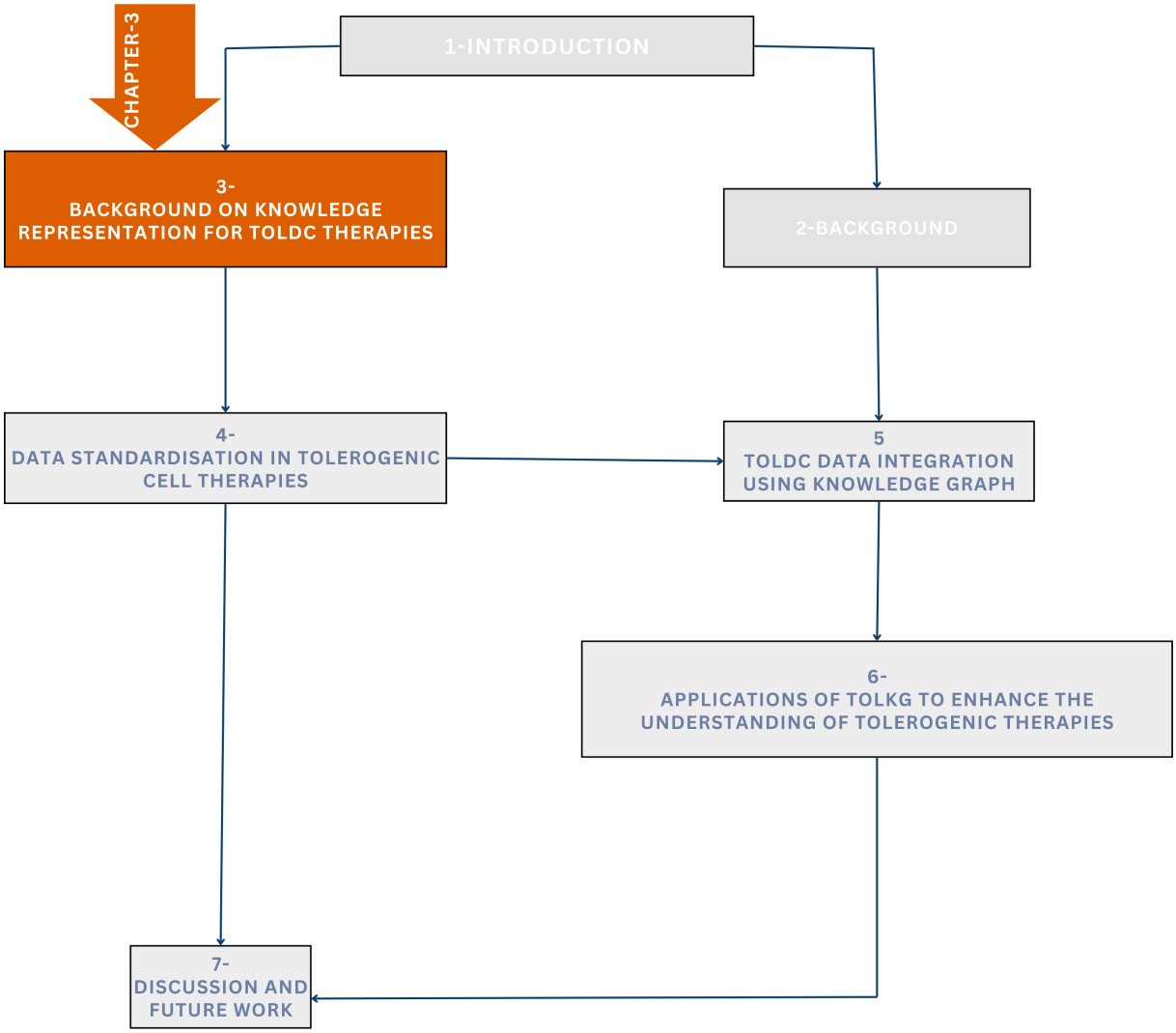


FIGURE 3.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

3.1 tolDC therapies

tolDC therapies represent a revolutionary direction in the field of immunotherapy, offering a potential solution to various autoimmune and inflammatory diseases [Suuring and Moreau, 2021]. DCs are vital components of the immune system, often referred to as the "conductors of the immune orchestra", as they play a crucial role in initiating and regulating immune responses. DCs can be programmed to induce either an immunogenic or tolerogenic response, making them a valuable tool in immunotherapy [Marin-Gallen et al., 2010].

DCs were first discovered by Jacques Banchereau & Ralph M. Steinman in 1973 [Banchereau and Steinman, 1998]. DCs promote immune tolerance by suppressing or moderating immune responses against self-antigens, thus preventing unwanted inflammation and autoimmunity. Through a combination of advanced genetic engineering and molecular biology techniques, it is now possible to generate tolDCs in a laboratory setting. These tolDCs can be used to create therapies that introduce tolerance to specific antigens, which has promising implications for the treatment of autoimmune diseases. Over the recent years, a significant knowledge expansion has happened in the field tolDCs. Several phase 1 clinical trials have also been conducted using tolDC-based immunotherapies such as type 1 diabetes, rheumatoid arthritis and multiple sclerosis [Giannoukakis et al., 2011, Kurochkina et al., 2018, Benham et al., 2015, Bell et al., 2017, Zubizarreta et al., 2019, Willekens et al., 2019]. Some studies also confirmed the applicability of tolDCs for enhancing transplantation survival rate [Morelli and Thomson, 2007, Turnquist et al., 2007, Moreau et al., 2012].

The protocols to generate these tolDCs consist of many steps and variations. Firstly, the tolDC can be generated using different types of agents such as Vitamin-D3 [Canning et al., 2001], rapamycin [Falcón-Beas et al., 2019], dexamethasone [Stenger et al., 2014] or a combination of these. Briefly, the protocol starts with extracting substrate cells from autologous or allogenic umbilical cord tissues by density gradient centrifugation of whole blood, bone marrow aspirate and digested tissue (lipo-aspirate and umbilical cord tissue) or by leukapheresis (whole blood). The cells are then cultured using the appropriate culture media, culture conditions and duration etc., which varies from protocol to protocol. These cultured cells can be either administered immediately or cryopreserved for administration in the patients later [Mosanya and Isaacs, 2019].

The general steps to generate tolDCs typically involve:

- Monocyte Isolation: Peripheral blood mononuclear cells (PBMCs) are isolated from blood samples, often through density gradient centrifugation. CD14+ monocytes are then selected from PBMCs using magnetic bead separation.
- Cell Culture and Differentiation: The isolated CD14+ monocytes are cultured in a medium containing cytokines that promote dendritic cell differentiation, typically interleukin-4 (IL-4) and granulocyte-macrophage colony-stimulating factor (GM-CSF).
- Induction of Tolerogenic Phenotype: To convert the differentiating dendritic cells into a tolerogenic state, various agents can be added to the culture. Common agents include vitamin D3, to promote a tolerogenic phenotype and immunosuppressive drugs like dexamethasone or rapamycin. Genetic modifications may also be employed to enhance the tolerogenic properties.
- Maturation (Optional): In some protocols, an additional step is included to induce maturation of tolDCs, using agents like tumour necrosis factor-alpha (TNF-α), although this may vary depending on the desired tolerogenic properties and the specific application of the tolDCs.
- Harvesting and Characterization: After the cultivation period, tolDCs are harvested and characterized to confirm their phenotype and functionality. This often involves assessing surface marker expression via flow cytometry and evaluating their capacity to stimulate or suppress T-cell responses.
- Application: Finally, the generated tolDCs can be used for various applications, including in vitro studies to understand mechanisms of tolerance, or in vivo for therapeutic purposes such as treating autoimmune diseases or in transplant tolerance protocols.

3.2 Data representation and integration challenges in tolDC therapies

This section highlights some of the major challenges for data representation in the field of tolDCs such as the field is new. As shown in Figure 3.2, yearly publications in the field of tolDCs are significantly smaller than the T cells, which are a broader category of immune cells and have been extensively studied for decades. The field of T cells is more mature, with a vast body of research dedicated to understanding their diverse roles in immunity, autoimmunity and cancer. Similarly, other related fields such as B cells or innate immune cells also have a more substantial research base. However, the field of tolDCs is growing; however, its smaller size does not diminish its potential impact on our understanding of immune regulation and the development of novel immunotherapies.

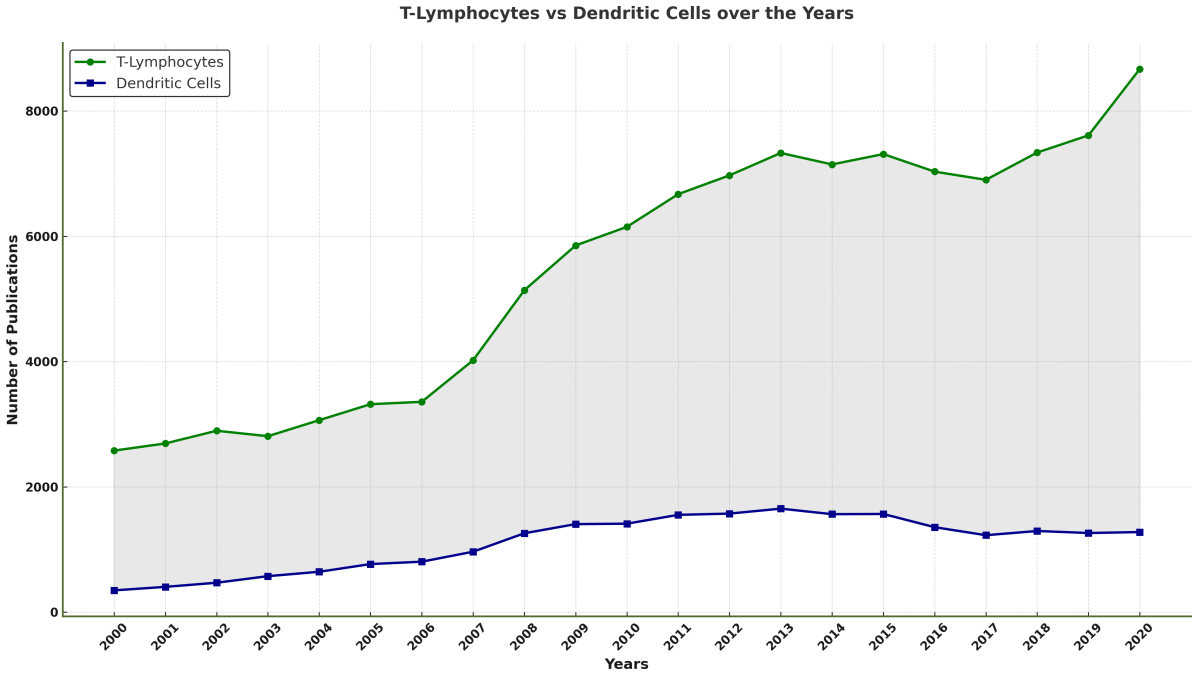


Figure 3.2: Graph showing the relative sizes of the tolDC vs T-Lymphocyte fields Data on tolDC was generated from TolKG (see Chapter 5) while T-lymphocyte papers were selected using MESH keywords from PubMed

3.2.1 The gap between research and associated data

The lack of comprehensive data reporting poses a significant challenge in the field of tolDC therapies. Despite the promising potential of tolerogenic approaches, the field suffers from incomplete or inadequate reporting of research findings and data, which hinders the advancement and translation of these therapies [Lord et al., 2016, Ioannidis, 2005, Boulton, 2016].

Figure 3.3 demonstrates a notably low number of research papers on "Dendritic cells" that also publish their experimental data alongside their research findings. This observation is based on all research conducted on dendritic cells and is not restricted to those focusing on tolerogenic DCs, where the rate is even lower. Interestingly, there was a substantial rise in data reporting after 2012. Nevertheless, as shown in Figure 3.3, we can see a significant drop in 2019 which was the start of the COVID-19 Pandemic and the situation has remained suboptimal since then.

Another aspect contributing to the lack of data reporting in the field is the early stage of development and clinical implementation of tolDC therapies. Many of the therapeutic strategies are still being explored in preclinical studies or early-phase clinical trials. Researchers may prioritise initial safety and efficacy evaluations, often leading to incomplete reporting of detailed data. Moreover, some studies may be terminated prematurely or not published due to negative or inconclusive results, creating a bias towards positive findings and further limiting the availability of comprehensive data.

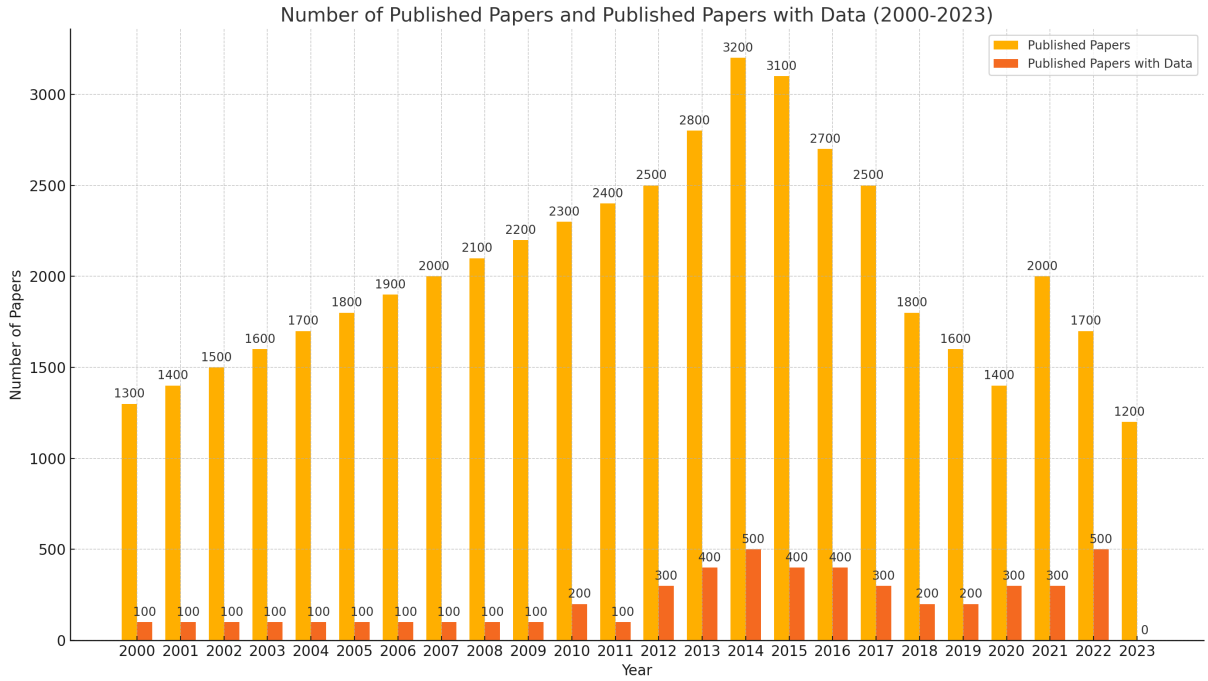


Figure 3.3: A comparison graph representing the number of published research papers with data vs without data. This figure shows data for all dendritic cell publications regardless if they are tolerogenic or not. Dated 18 June 2023

3.2.2 Unavailability of a specific datawarehouse

Databases play a critical role in consolidating and organising data in a structured and accessible manner, thus enabling researchers to find information efficiently. However, the field of tolDC therapies, despite its potential to provide treatment for various immune-related conditions, has been operating without a specialised database which hinders research progress. For instance, researchers often struggle to locate the information they need without a centralised database which makes it difficult to build on previous findings, confirm results or investigate new hypotheses [Dall'Olio et al., 2010].

This situation significantly contrasts with other research fields where dedicated databases exist. For example, genomics research has hugely benefited from databases like GenBank, which stores annotated collections of all publicly available DNA sequences [Res, 2012]. Similarly, the Protein Data Bank (PDB) serves as a worldwide repository of information about the 3D structures of proteins, nucleic acids and complex assemblies [Berman et al., 2000]. Such databases are helpful in the research process and expediting scientific discoveries. Therefore, the development of a dedicated database for tolDC therapies could significantly benefit this field. It can stimulate the efficient sharing of experimental data, facilitate the replication of research and accelerate the pace of discovery and innovation.

In the absence of a dedicated online database for tolDC therapies, a thorough review of the literature was undertaken to identify potential relevant databases related to tolDCs. The most pertinent databases identified include LinkedImm [Bukhari et al., 2019], ImmPort [Bhattacharya et al., 2018], ImmuneData [Deng et al., 2022] and ImmGen [imm, 2020], each specialising in immunology. Given that tolDC therapies fall within the field of immunology, these databases are anticipated to contain relevant tolDC data. However, as shown in Table 3.1, a surprising observation is the scarce or even non-existent data entries on tolDC within these databases. The most substantial collections of tolDCs data were found in the GEO (Gene Expression Omnibus) [Barrett et al., 2012] and ArrayExpress databases [Athar et al., 2019]. When conducting these searches, no restrictions on the date or data type were imposed, with the intention of obtaining the most comprehensive results possible.

Table 3.1: Availability of tolDC related data at different relevant databases

Database	**Link**	**Entries on tolDC data**
LinkedImm	https://linkedimm.org	None
ImmPort	https://www.immport.org	8
ImmuneData	https://immunedata.org/	None
ImmGen	https://www.immgen.org	10
GEO	https://www.ncbi.nlm.nih.gov/geo	80
IEDB	https://www.iedb.org	None
TCGA	https://portal.gdc.cancer.gov	None
ArrayExpress	https://www.ebi.ac.uk/biostudies/arrayexpress	53

3.2.3 Lack of data standardisation

The lack of standardised data collection and reporting in the field of tolDC therapies poses another significant challenge to researchers and clinicians. This problem arises due to the diverse nature of tolerogenic strategies, variations in study designs and the absence of a consensus on reporting guidelines [Ten Brinke et al., 2015].

One of the primary issues related to data collection is the heterogeneity in experimental protocols and methodologies employed in different studies. tolDC therapies encompass a range of approaches, including antigen-specific tolerance induction, regulatory T-cell modulation, immune checkpoint manipulation and more. Each approach may involve distinct procedures, dosing regimens and evaluation criteria. This heterogeneity makes it difficult to compare and integrate data across studies, hindering the development of a standardised framework for assessing the effectiveness and safety of these therapies. Moreover, the selection of relevant biomarkers and outcome measures can vary across studies. Biomarkers play a crucial role in assessing the immune response and therapeutic outcomes in tolDC therapies. The choice of biomarkers can vary depending on the specific research objectives, the target disease and the stage of therapy development. Without standardised biomarker selection and reporting, it becomes challenging to establish a consistent and comprehensive dataset to evaluate the efficacy and long-term effects of tolerogenic interventions [Hilkens and Isaacs, 2013].

Additionally, the lack of standardised data reporting practices leads to incomplete and inconsistent data presentations. Researchers may selectively report certain outcomes or focus only on statistically significant results, which can introduce reporting bias. Incomplete reporting can make it challenging to assess the robustness of the research findings, replicate studies and perform comprehensive meta-analyses [Gosselin, 2021, Grigorian-Shamagian et al., 2021]. The absence of detailed data on study design, patient characteristics, treatment protocols and adverse events further obstructs the evaluation and comparison of different therapeutic approaches.

In summary, the lack of standardised data collection and reporting in the field of tolDC therapies hampers the progress, comparability and reproducibility of research findings.

3.2.4 Data heterogeneity and complexity

Data heterogeneity and complexity also pose significant challenges in the field of tolDC therapies. The intricate nature of the immune system and the diverse range of approaches employed in tolerogenic interventions contribute to the variability and complexity of data generated in this field. One key aspect of data heterogeneity is the variation in study designs and experimental protocols. This diversity in experimental design leads to differences in data generation, making it challenging to compare and integrate findings across studies.

Moreover, data heterogeneity arises from the use of various analytical methods and technologies across studies. Different research groups may employ different assays, platforms, or techniques to measure immune responses or biomarkers. These variations can introduce discrepancies in the data and make it difficult to compare results directly. Standardising the assays and analytical methods used for data collection can help reduce data heterogeneity and enhance comparability across studies. The complexity of immunological data further exacerbates the challenges in the field. The immune system is a highly intricate and dynamic network of cells, molecules and signalling pathways. Assessing the effects of tolDC therapies often requires the evaluation of multiple parameters, including immune cell phenotypes, cytokine profiles, functional assays and molecular signalling pathways. The collection and analysis of such complex datasets require specialised techniques, expertise and sophisticated bioinformatics tools.

Addressing the challenges of data heterogeneity and complexity requires collaborative efforts and standardisation initiatives. Harmonising study designs, protocols and data collection methods can promote consistency and comparability across different research groups. The development and adoption of standardised assays, protocols and quality control measures for immune profiling can enhance the reliability and reproducibility of data. Furthermore, integrating multi-omics data and systems biology approaches can provide a more comprehensive understanding of the complex immune response in tolDC therapies. By combining transcriptomic, proteomic and metabolomic data, researchers can gain insights into the molecular mechanisms underlying therapeutic effects and identify potential biomarkers of response or resistance. The advancement of computational tools and bioinformatics approaches is also crucial in managing the complexity of immunological data. Machine learning algorithms, network analysis and data visualisation techniques can aid

in the identification of patterns, clustering of patient subgroups and prediction of treatment outcomes. These computational approaches can help uncover meaningful insights from complex datasets and facilitate the translation of data into actionable knowledge.

In addition to the above, several different kinds of technologies are being used to analyse the tolDCs including flow cytometry, ELISA (Enzyme-Linked Immunosorbent Assay), Western blotting, polymerase chain reaction (PCR) and next-generation sequencing (NGS), among others [Maecker et al., 2012, Toussi and Massari, 2014, Schultze and Aschenbrenner, 2021]. However, each of these techniques carries its inherent characteristics, which may lead to data heterogeneity. The table below explains how these different techniques contribute to data heterogeneity and complexity.

Technique	Description	Source of Heterogeneity
Flow Cytometry	Primarily used for the analysis and sorting of individual cells based on their physical and chemical characteristics. It identifies different cell populations, such as regulatory T cells or antigen-presenting cells.	Differences in flow cytometer configurations, staining protocols and data analysis techniques.
ELISA	Used to detect and measure specific proteins or antibodies in a sample. It could be employed to measure cytokines or other immunoregulatory molecules.	Inter-laboratory differences in protocols, variations in antibody batches and differences in sample preparation.
Western Blotting	Detects specific proteins in a sample and can be used to study protein expression related to immune tolerance.	Variations in protocol, such as antibody selection and use, as well as differences in gel electrophoresis conditions.
PCR	Employed for amplifying a specific DNA sequence, often used to study gene expression related to immune tolerance.	Different protocols, primers and conditions.
Next-Generation Sequencing (NGS)	High-throughput sequencing method that allows for the sequencing of entire genomes or the targeted sequencing of particular areas of interest. Could be used to identify genetic factors contributing to immune tolerance.	Variations in library preparation, sequencing platforms and bioinformatic analysis.

Table 3.2: Sources of heterogeneity in different data analysis techniques

3.3 Text mining for knowledge extraction

As discussed, there is a lack of experimental data availability in the tolDC field. Although the number of publications is less than in other fields but there is still literature available on the tolDC field, which contains knowledge about the field. If we want to extract this knowledge from the literature, it is unrealistic to keep up with the amount of data published in scientific literature through manual curation. The sheer volume of information, especially in biomedical fields, makes it even more challenging to process and understand without the assistance of technology. In this case, automated approaches for the extraction of data, whether it is specific biomedical terms or intricate inter-relations between them, offer a significant advantage. These automated systems use techniques like text mining to analyse the literature, unlocking information that would otherwise remain hidden in an overwhelming set of free-text documents.

Text mining in biological fields has its origins in the late 1990s and early 2000s [Chen and Sharp, 2004, Hirschman et al., 2005]. Early applications were focused on tasks such as gene and protein name recognition within the burgeoning field of genomics [Hunter and Cohen, 2006]. The challenges of handling specific biological terminologies and the interconnection between various biological concepts led to the development of specialised text mining tools tailored to these fields [Bodenreider, 2006].

The potential to use literature mining for tolDC therapies is huge. Text mining can search through tons of documents to find exactly what scientists need, saving time and effort. It can pinpoint the latest research, discoveries and methods related to tolDC therapies. In tolDC therapies, many things are interconnected, like genes, proteins and diseases. Text mining can help map these connections by analysing scientific texts and helping researchers see how everything fits together.

3.4 Standardisation approaches

We discussed different examples of data representation at different time periods in Section 2.1, here we briefly discuss about the standardisation approaches. In the 20th century, the development of molecular biology and genetics led to the need for standardised terminology to describe genes, proteins and genetic sequences. Organisations like the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of

Biochemistry and Molecular Biology (IUBMB) have played roles in standardising chemical and biochemical nomenclature. With the advent of large-scale data collection in biology, the need for standardised vocabularies and ontologies has became critical. The Gene Ontology (GO) and other ontology projects were established to provide standardised terms and relationships for describing gene functions, biological processes and cellular components [Edgar et al., 2002]. Similarly, biological databases and repositories such as array express use standardised metadata and terminologies to ensure data interoperability [Brazma et al., 2003]. The Human Genome Project and subsequent genomics projects further accelerated the need for standardised terminology and data formats. Initiatives like the Minimum Information About a Genome Sequence (MIGS) and Minimum Information About a Microarray Experiment (MIAME) guidelines were developed to ensure consistent reporting of genomic and transcriptomic data [Field et al., 2008, Brazma et al., 2001].

In various scientific and technical domains, minimal information models (MIMs) are implemented to standardise the reporting of data and information, enabling consistent communication, comparison and integration across different studies or systems. MIMs are reporting frameworks that describe the essential information that needs to be provided in a publication so that the work can be repeated or compared to other work [Snickars and Weibull, 1977]. MIMs consist of a checklist of information for experimental data reporting and are created by experts in the field. This checklist is divided into *MUST, SHOULD, MAY* requirements. One of the main motivations behind defining a checklist is to ensure that the reported metadata is sufficient for independently interpreting the data, thus enhancing its reusability and enabling open data research. If the MIM is not followed, providing complete information about the experiment can be difficult and the level of detail will differ by different researchers. There has been a growing interest and stress in following the MIMs in all fields of life sciences, as can be seen by the increasing number of available MIMs as shown in Figure 3.4.

In many experiments, relevant information is not contained in the experimental data file itself. As the experiment becomes more complex and diverse, it has become crucial and challenging to report the method fully. A solution to resolve this issue was identified in the form of a MIM for tolDCs [Lord et al., 2016] named MITAP. The primary purpose of MITAP was to transparently explain all the steps followed in the experiment including the reagents, data processing and findings of the experiment. This metadata or contextual

data is often important in the downstream analysis of the experiment [Hippen and Greene, 2021] because metadata is crucial for reproducible research. There has been a growing interest and effort to promote metadata sharing overall in the biomedical field. Moreover, more journals and data-sharing platforms require researchers to follow reporting guidelines or frameworks.

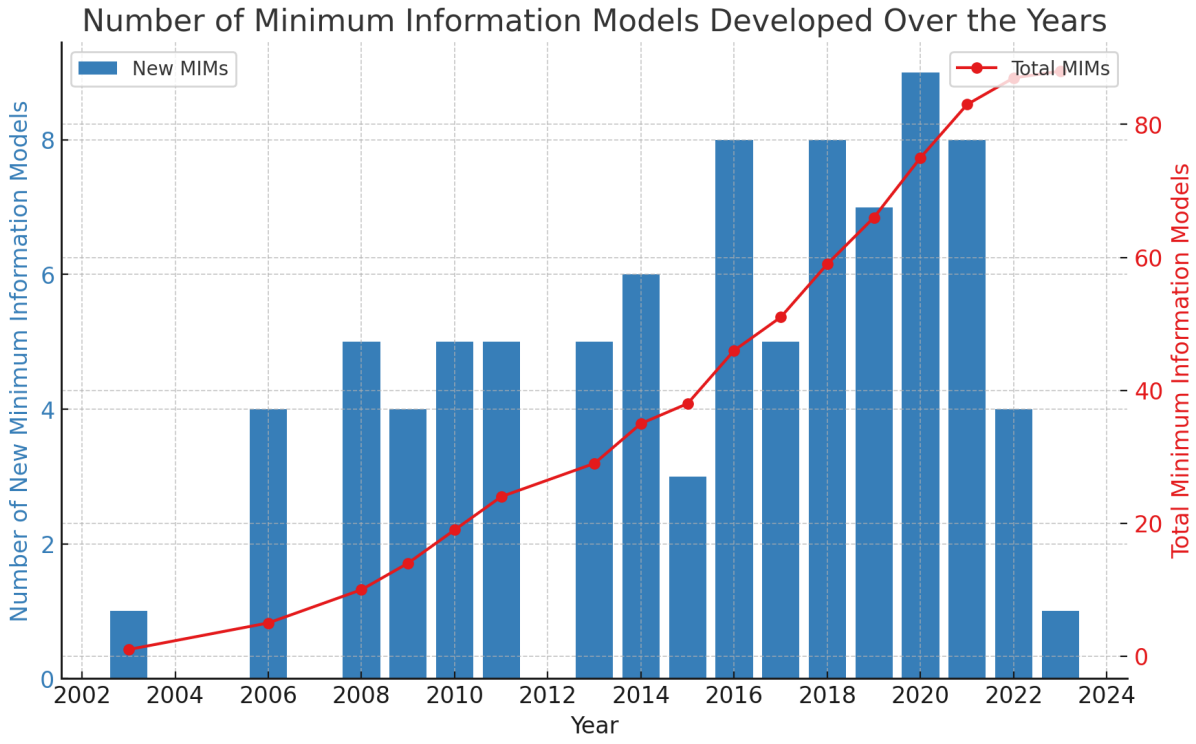


Figure 3.4: Graph representing the number of minimum information developed over the years. The data in this graph is generated using the data from PubMed by putting the filter that the title of the publications should have "Minimum Information Model" and then refined by manual inspection.

3.5 Promoting MITAP and tolDC Knowledge Graph

Raising awareness of tolKG and MITAP among researchers, clinicians, and industry professionals is essential for adoption. Workshops and training sessions tailored to immunologists, bioinformaticians and clinicians can introduce tolKG and demonstrate its role in research and hypothesis generation. Training on MITAP compliance will help researchers align experimental protocols and data reporting with its standards. Hands-on tutorials can guide users in querying tolKG and structuring data accordingly.

Collaboration with research institutions and universities can integrate tolKG and MITAP into research workflows. This includes partnering with immunology and bioinformatics groups to incorporate tolKG into data analysis pipelines, encouraging its inclusion in grant proposals, and embedding it in postgraduate training programs. Clinicians using tolDC-based immunotherapies can benefit from structured data reporting, positioning tolKG as a clinical decision-support tool and MITAP as a standard for trial documentation. Engagement with regulatory bodies (e.g., EMA, FDA) will align MITAP with existing reporting standards for cell-based therapies.

Integration with Biomedical Databases Enhancing tolKG and MITAP's visibility requires integration with key biomedical databases. tolKG should be linked to NCBI Gene, GEO, ArrayExpress, ImmPort, IEDB and LinkedImm to ensure interoperability and increase its utility for immunology research.

Encouraging MITAP Data Deposition Researchers should be incentivized to deposit tolDC-related datasets in public repositories using MITAP-compliant metadata. This can be facilitated by providing MITAP templates for data submission, developing automated compliance-checking pipelines and encouraging journals and funding agencies to mandate MITAP adherence.

Dissemination via Conferences and Publications Presenting tolKG and MITAP at major conferences in immunology, bioinformatics and data science will boost adoption. Publishing in high-impact journals will further establish credibility and increase engagement within the scientific community.

Community-Driven Initiatives Building a strong user community will ensure the long-term adoption of tolKG and MITAP. Establishing an online forum, Slack or Discord channels, and hosting webinars will foster collaboration. Researchers should be encouraged to contribute by annotating new tolDC-related publications with standardized MITAP metadata and manually curating relationships in tolKG.

Aligning tolKG and MITAP with FAIR (Findable, Accessible, Interoperable, Reusable) principles will enhance their impact. Promoting MITAP-compliant datasets as FAIR-compliant models and integrating tolKG with semantic web technologies will improve accessibility and usability.

3.6 Summary

Recent advancements in the generation and application of tolDCs in clinical settings highlight the potential of tolDC-based therapeutic approaches. The ability to tailor these tolerogenic cells in vitro to address specific antigens offers an efficient therapeutic option that aligns with the principles of precision medicine. Moreover, the ongoing clinical trials and studies exploring tolDCs in various contexts are promising indicators of the broader applicability and effectiveness of this therapy.

However, as the field evolves, it is crucial to address the challenges associated with data representation and integration. The development of a dedicated database for tolDC therapies, alongside standardised data collection and reporting protocols, is crucial in overcoming these obstacles. Such advancements will not only enhance our understanding of tolDCs but also accelerate the translation of research findings into clinical practice. As discussed in this chapter, the tolDC field can benefit and progress rapidly by fostering an environment of comprehensive data sharing and collaboration.

4

Data Standardisation in tolerogenic dendritic cell therapies

4.1 Abstract

In this Chapter, we consider how standardisation has impacted the tolDC field. Specifically, we consider the uptake of the reporting standard, Minimum Information about Tolergenic Antigen-Presenting Cells (MITAP), which was introduced in 2016. We also consider whether the use of MITAP affects the structure and availability of metadata presented in papers that have used it. We have applied literature analysis of representative papers, which was then used to generate queries against PubMed, allowing us to build a corpus of the entire tolDC field; finally, a sample of these papers was manually analysed by experts looking at the use of metadata. As a result of this, we conclude that MITAP is being used and has increased the amount of metadata available, but at relatively low levels.

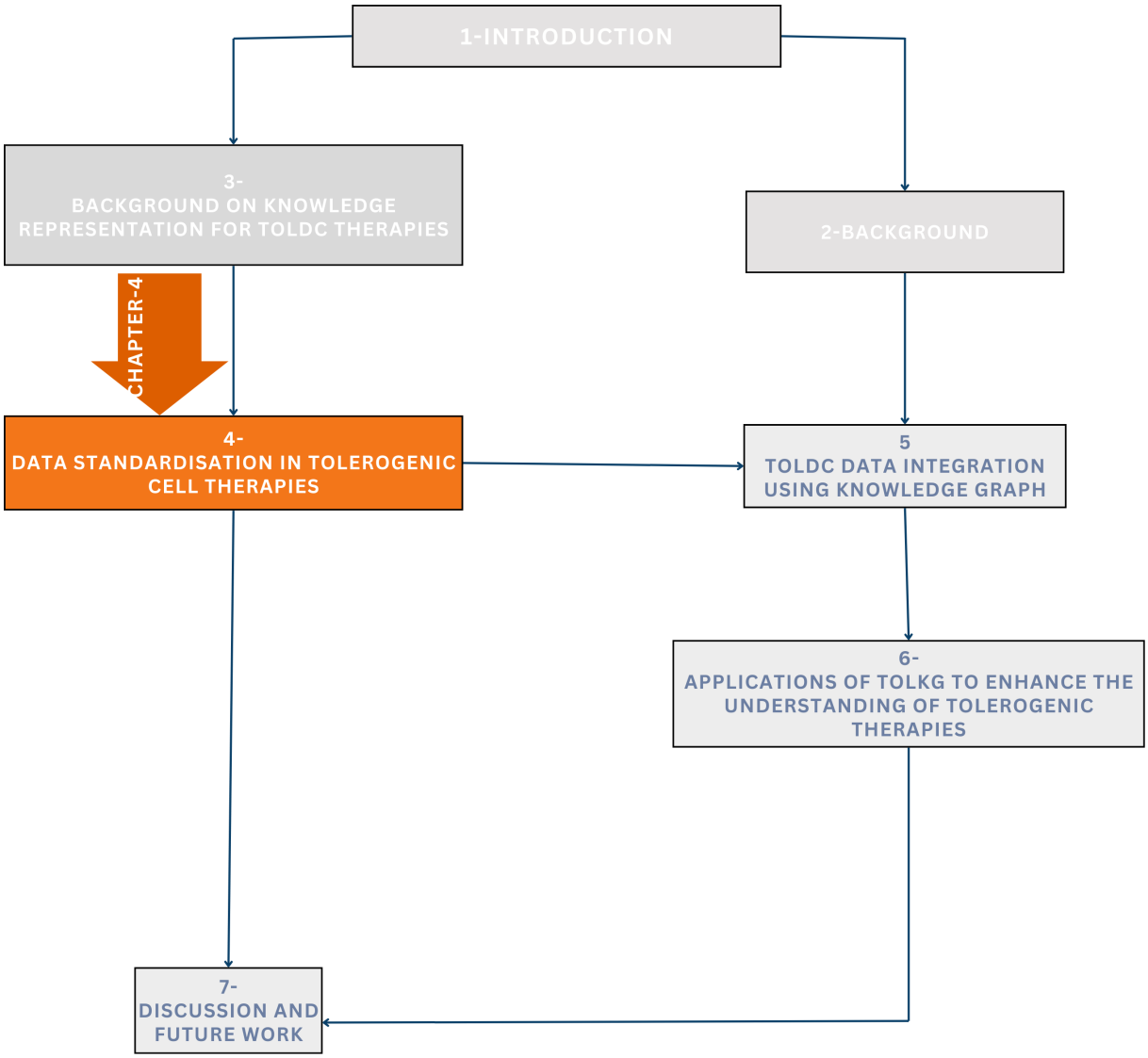


Figure 4.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

4.2 History

This chapter is based on the publication "Tolerogenic Dendritic Cell Reporting: has MITAP made a difference?"[Sahar et al., 2023]. The data in this chapter is also presented in that paper. The introduction is similar but partially rewritten to fit this thesis.

4.3 Introduction

As discussed in Chapter 2, it is difficult to compare different tolDC protocols because of the variations across different research experiments. To partially resolve this issue, experts introduced a solution called the Minimum Information about Tolerogenic Antigen-Presenting Cells (MITAP) in 2016 [Lord et al., 2016]. This is a type of MIM for tolDCs, and its goal is to help standardise the reporting of metadata. Now, almost five years after publishing it, it is important to check how much impact MITAP has had on the tolDC field.

The first step to achieve this is to evaluate whether the use of MITAP has led to a significant improvement in metadata reporting compared to the scenarios when it is not used. Moreover, the extent of adoption of MITAP by the research community needs to be assessed, since it will not be effective if it is not used. In other words, the adoption of MIMs is pivotal in promoting reproducibility, data sharing and integration. Several factors may act as barriers to the adoption of MIMs: one is the burden of compliance, as researchers may see the process of documenting experiments as time-consuming; another is the lack of incentives or requirements for compliance, which may discourage researchers from adopting MIMs that require excessive effort.

One way to encourage the adoption of MIMs is to incorporate them into research funding policies and requirements. Funding agencies can mandate that researchers adhere to specific reporting guidelines for their data if they are to receive funding. Additionally, journals can play a role in promoting the adoption of MIMs by requiring authors to comply with reporting guidelines as a condition of publication. These approaches can enhance their awareness and adoption among researchers, promoting consistency and standardisation in the published literature. Researchers may not recognise the potential benefits of standardised reporting or may not fully understand how to implement the guidelines in their research.

To overcome these challenges, several strategies can be employed to increase the adoption of MIMs. One such approach is to provide education and training to researchers about the importance and benefits of standardised reporting, as well as practical guidance on implementation in their research. In conclusion, while MIMs can be a powerful tool for promoting reproducibility and facilitating data sharing and integration in scientific research, their effectiveness ultimately depends on their adoption by the research community. Addressing barriers to adoption and promoting awareness and education about the importance and benefits of standardised reporting can increase the uptake of MIMs and improve the reliability and impact of scientific research which is the focus of this chapter.

In this chapter metadata reporting and MITAP use are discussed along with the challenges faced in following the standardisation protocols. A literature-based method to find out all the publications that followed MITAP is explained. A comparison between following the MIM in tolDC field with other fields is also presented. This analysis shows that MITAP does increase the quantity and organisation of metadata that is presented in papers, therefore highlighting their importance to the field. It concludes by discussing ways to make MIMs easier to follow.

4.4 Background

A large amount of background information is required to fully understand the context, methods, data and conclusions that pertain to an experiment. MIMs provide a framework for reporting all the essential information (metadata) about an experiment and they have become popular among the scientific community [Taylor et al., 2008]. MIMs are developed keeping in mind that they do not complicate the experimental process nor become a burden on the community, hence "minimum". Moreover, MIMs are designed to make research more reproducible; following them ensures that the necessary information about the experimental work is presented in a report or publication so that other researchers can reuse or repurpose the data and methods.

tolDCs are antigen-presenting cells that can induce or restore immune tolerance [Fucikova et al., 2019]. These cells show great promise as a therapeutic tool for the treatment of conditions caused by a breach in immune tolerance (e.g. autoimmune diseases) or for the

prevention of graft rejection. However, the diverse range of experimental designs and reagents used to generate tolDCs preclude meaningful comparisons between different tolDC types. It is the wide variation between protocols that encouraged a group of researchers working in the field (of tolDCs) to come together between 2014-16 to generate a MIM for their field.

4.4.1 Minimum Information Model for Myeloid Regulatory Cell Therapies (MITAP)

In 2016, MITAP was created to standardise the reporting of tolerogenic antigen-presenting cells, including tolDCs. It was anticipated that a MIM would improve the transparency, reproducibility and data interpretation of the different tolDC types. Its more general name was chosen to allow for the inclusion of other tolerogenic antigen-presenting cell types, for example, regulatory macrophages (MRegs).

MITAP was collaboratively created by experts in the tolDC field over a period of 18 months. The whole procedure involved a series of interactive workshops in which several exercises were conducted to gather the relevant basic vocabulary used within the community both to obtain feedback on the draft reporting guidelines and finally to test its comprehensibility by the end-users. The full MITAP document, including the checklist, can be found on http://w3id.org/ontolink/mitap.

MITAP consists of 4 sections that were considered to be essential by the experts for reporting all necessary information about the generation of tolDCs and other tolerogenic antigen-presenting cells (tolAPC) products. The most crucial stages of the production process are encapsulated in these sections in an orderly manner. They are summarized below:

Section 1. Cells before:

This section describes the characteristics of the cells before they undergo any manipulation such as (a) essential information about the donor, (b) source of the cells, (c) the cell extraction method (d), cell phenotype after extraction and (e) cell number and viability.

Section 2. Differentiation and induction of tolerogenicity:

This section describes the protocol that has been used to differentiate and/or induce tolerogenicity in the cells described in Section 1. There are five subsections giving details

on (a) preculture conditions, (b) culture conditions, (c) the protocol used to induce differentiation and tolerogenicity of the cells, (d) loading of cells antigen and (e) the way cells are stored immediately after culture.

Section 3. Cells after:

This section describes the characteristics and state of the cells after the differentiation/induction of tolerogenicity process has taken place. This section has three subsections, two of which provide similar basic details as in Section 1 on the (a) cell phenotype and (c) cell number and viability. Another part (b) focuses on functional in vitro assays such as the migratory capacity of the cells, or their ability to induce T regulatory cells.

Section 4. About the protocol:

The final section of MITAP describes the general features of the protocol as a whole, such as a) any external regulatory authorities or guidelines followed; b) the purpose of the cells; c) whether cell product is applied in an autologous, allogeneic, xenogeneic or syngeneic manner and finally d) contact details of the authors.

4.5 Investigating the impact of MITAP on tolDC field

MITAP was published to improve the data reporting in the tolDC field. We wanted to investigate the impact of MITAP in the field since it was published almost 4 years before the time of this analysis. The main aim is to investigate the usage of MITAP by researchers who are publishing immunology papers. More specifically, we addressed the question of whether these reporting guidelines improved the provision of relevant metadata in papers published after the MITAP publication date. To do this we need to find papers that are generating tolDCs regardless of whether they are using MITAP which is not a simple task; from this set, we also need to find those papers that are using MITAP. We describe the methodology that we have used to achieve this here. Finally, the comparison between the data reporting of publications that used MITAP is compared to those that did not.

4.5.1 How many papers have used MITAP?

The first stage of our analysis was to ask the question of how many papers used MITAP to organise and characterise their datasets. To answer this, we assumed that any paper that

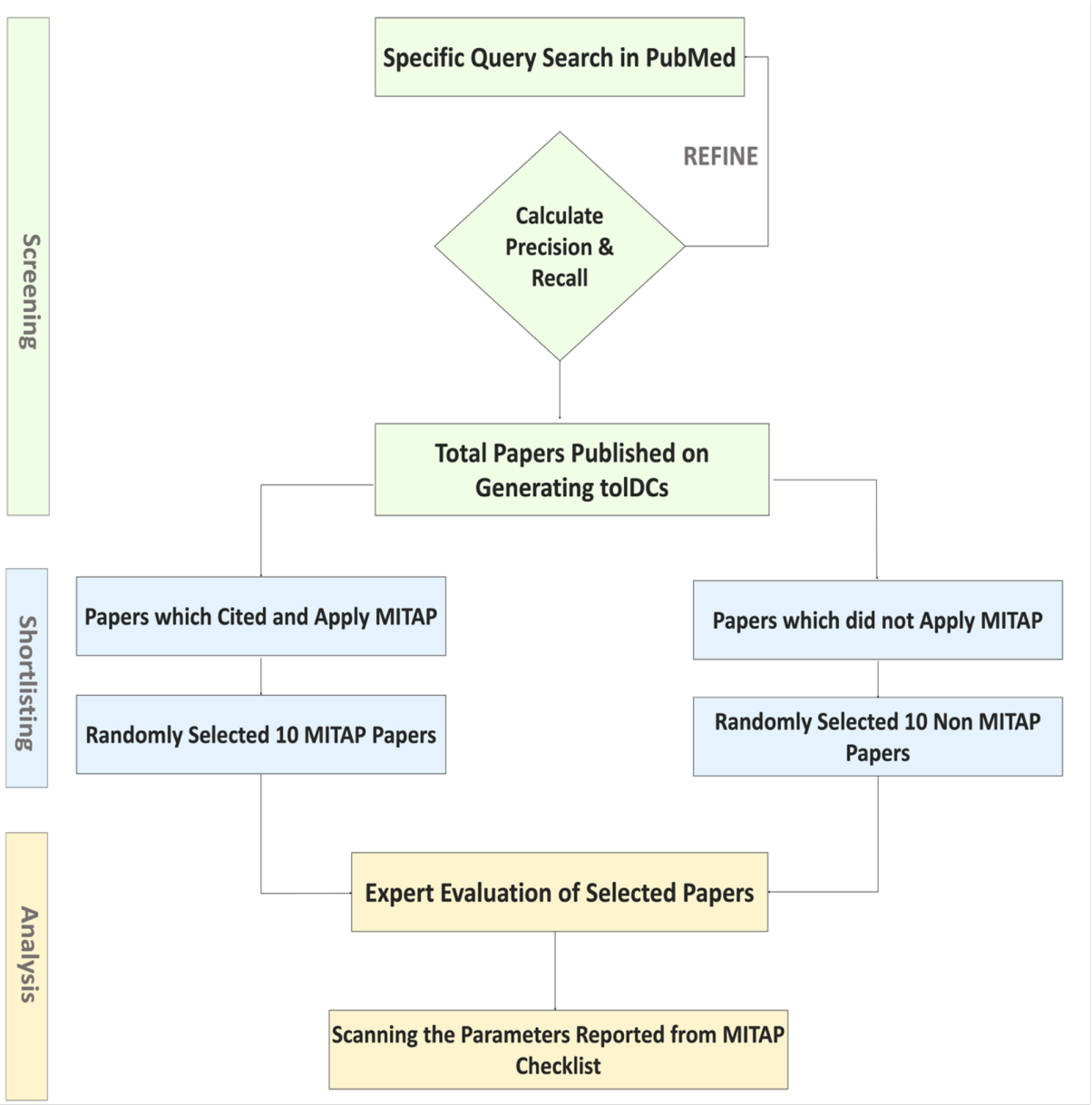


Figure 4.2: Overall approach for Investigating the Impact of MITAP on tolDC field. The above diagram represents the overall approach for Investigating the Impact of MITAP on tolDC field. The process begins with a specific SQL query-based search on PubMed, with further refining the search by using the most relevant keywords, and a corpus of all the papers published on generating tolDCs is established. An unbiased random selection of 10 papers is performed from each category and these papers are further investigated by experts in the field of tolDCs.

directly used MITAP would include it in the reference list. Therefore, we used Google Scholar and downloaded all the papers that had cited MITAP; we found that MITAP had 40 citations in January 2021. By inspection, we found that not all of these papers directly used MITAP; some papers, for example reviews, referred to MITAP as an example of a MIM. In fact, only 10 out of the 40 papers that had cited MITAP were using it directly as a reporting framework.

4.5.2 How many papers could have used MITAP?

While it is useful to know the actual usage of MITAP, we also wished to understand how many papers that are reporting results about tolDCs could have used MITAP but did not. This question is harder to answer exactly, but we calculated an estimate. Our overall approach is depicted in Figure 4.2. We assumed that most of the papers related to tolDCs are available on PubMed. We therefore searched PubMed with a variety of keywords appropriate for papers related to the MITAP subject matter, restricting the search to papers that were published after MITAP itself was published.

4.5.2.1 Building a relevant corpus

Of course, to use the number of search results as an estimate for the total number of papers, we must be sure of our query terms. We achieved this by testing the performance of each query by calculating the precision (how accurate the results were) and recall (how complete the results were). We could test for recall because we had a set of papers that directly cited MITAP; these therefore fall into the category of papers that could (and in this case did) use MITAP. If our query has 100% recall, therefore, all of these papers (and others) should be retrieved. We calculated the precision by simply reading a subset of 20 papers from each query and making a judgement about whether they were relevant; this analysis was performed by the author and two other experts who have degree-level or above expertise in the tolDC field.

Each researcher reviewed seven to eight papers to ensure the workload was evenly distributed and to prevent excessive burden on any individual. The researchers held regular meetings to discuss discrepancies and refine their screening criteria, ensuring that a consistent level of filtration was applied across all assessments. These discussions were particularly useful in resolving borderline cases, where some relevances were unclear and

in refining the query terms to optimise precision and recall for subsequent searches.

4.5.2.2 Term extraction and term frequency-inverse document frequency (TF-IDF)

TF-IDF is a statistical method to rank the critical terms in a document or a set of documents. It is extensively used in text mining for term extraction. The result of TF-IDF is a measure that increases as the frequency of a specific term increases in a document (referring to as TF) and decreases if the term appears more in the corpus (referring to IDF). This method helps to find words that are important in one document but common in many, making it easier to identify key terms that distinguish different documents. By reducing the influence of frequently used but less useful words, TF-IDF improves the performance of search engines and other language processing tasks.

t= Any term in the document d= document D= Corpus of documents

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \times \text{IDF}(t, D)$$

where:

$$\text{TF}(t, d) = \frac{\text{Number of times term } t \text{ occurs in document } d}{\text{Total number of terms in document } d}$$
$$\text{IDF}(t, D) = \log\left(\frac{\text{Total number of documents in corpus } D}{\text{Number of documents containing term } t}\right)$$

TF-IDF also filters out the common verbs or nouns such as and, etc. Thus, it is used to find the most important keywords in the titles and abstracts of the relevant corpus of tolDCs. These keywords are later used in the queries of PubMed. Hence, Table 4.1 represents the representative terms of the tolDC field.

4.5.2.3 PubMed query formation

To provide a baseline for our queries, we initially generated keywords using TF-IDF from the titles and abstracts of 10 relevant MITAP citations. All the queries were refined by putting a *date of publication* filter, so that only papers published after MITAP were retrieved, since no papers published before MITAP could have used it. In addition, the review papers were also excluded by applying a "NOT" filter on the query, since these

were not relevant to this investigation. A query using only keywords that were given in the original MITAP paper retrieved a relatively smaller number of papers, but the recall was also lower (see Table 4.1). Finally, we modified these queries based on the expert judgement of the authors resulting in the manually adjusted queries. The precision and recall were higher and the number of retrieved papers was large. Therefore, the query was further refined by adding keywords such as *generate*, *produce* or *induce* to get only those papers that reported on tolDCs that were produced as part of an experiment. The precision and recall of *manual adjustment-2* retrieved results were found appropriate among other queries. Thus, this list was decided to be the "total number of tolAPC papers published after MITAP" (of which the majority focused on tolDCs, as expected). Hence only 14% of papers that could have used MITAP to report their data have actually done so.

Table 4.1: Comparison of query methods for retrieving tolDC-related papers

Method	**Query**	**Number of Papers**	**Precision**	**Recall**
IDF via Titles	(("tolerogenic-dendritic"[Title] OR "derived-dendritic"[Title] OR "tolerogenic*"[Title] OR "regulatory-cell*"[Title] OR "regulatory-macrophage*"[Title]) NOT (review[Title/Abstract])) AND (("2016/08/30"[Date - Publication] : "3000"[Date - Publication]))	921	0.3	0.75
IDF via Abstracts	((("tolerogenic-dendritic"[Title] OR "derived-dendritic"[Title] OR "regulatory-cell*"[Title] OR "tolerogenic*" [Title] OR "regulatory-macrophage*"[Title]) OR (autoimmun*[Title])) OR (antigen-presenting[Title])) NOT (review [Title/Abstract])) AND (("2016/08/30"[Date - Publication] : "3000"[Date - Publication]))	8,992	0.25	0.71

Continued on next page

Table 4.1 – continued from previous page

Method	Query	Number of Papers	Precision	Recall
IDF via MITAP keywords	((Tolerogenic dendritic cell[Title] OR Regulatory dendritic cell[Title] OR Tolerogenic antigen-presenting cell[Title] OR Regulatory macrophage[Title] OR Tolerogenic dendritic cells[Title] OR Regulatory macrophages[Title]) NOT (review[Title/Abstract])) AND (("2016/08/30"[Date - Publication] : "3000"[Date - Publication]))	129	0.45	0.69
Manual adjustment version 1	(("tolerogenic-antigen-presenting"[Title] OR "tolerogenic-dendritic-cell*"[Title] OR "derived-dendritic*"[Title] OR "regulatory-macrophage*"[Title]) NOT (review[Title/Abstract])) AND (("2016/08/30"[Date - Publication] : "3000"[Date - Publication]))	410	0.5	0.69

Continued on next page

Table 4.1 – continued from previous page

Method	Query	Number of Papers	Precision	Recall
Manual adjustment version 2	(("antigen-presenting"[Title] OR "dendritic-cell*"[Title] OR "derived-dendritic*"[Title] OR "regulatory-macrophage*"[Title]) AND ("induc*"[Title] OR "generat*"[Title] OR "develop*"[Title] OR "produce*"[Title]) NOT (review[Title/Abstract]) AND ("tolerogenic*"[Title])) AND (("2016/08/30"[Date - Publication] : "3000"[Date - Publication]))	72	0.75	0.83

4.6 Comparison of papers with and without MITAP

The purpose of the MITAP document is to ensure that authors provide sufficient basic information about the generation of tolDCs or other types of tolAPC. It was routine that complete information was not provided. Five years after the MITAP publication, we wanted to re-check the status of information provided by the tolDC community. We therefore applied the MITAP checklist to 10 papers that used MITAP and 10 papers that could have used MITAP but did not. These papers are listed in the reference list chronologically, but the sequence in Figure 4.3 is different and anonymised. Red sections in the heatmap represent the information missing from the papers, green areas show the information is directly provided in the paper and yellow areas show that the information is partially available (e.g. information is available in a referenced paper but not in the paper itself), or not relevant to the paper. Figure 4.3 shows that not all the sections of the MITAP checklist were completed by both categories.

A clearer comparison is shown in Fig 4.4, which highlights the percentages demonstrating a significant difference in the reporting of Sections 1, 2, and 4. In contrast, Section 3 does not exhibit a significant difference between the papers that used MITAP and those that did not.

4.6.1 Top reported and unreported fields

We next wanted to understand the most and least reported fields among papers. The Figure 4.5 shows the most reported fields in MITAP papers and non-MITAP papers. More fields have been reported by all papers which followed MITAP than those that did not. In addition, some fields are reported by all MITAP and Non-MITAP papers such as cell surface molecules, methodology, name and contact details of the corresponding author.

MITAP checklist has 3 types of fields, **must**, **should** and **may**. The fields are highlighted according to these types.

Moving on to the top unreported fields in MITAP vs Non-MITAP, the number of top unreported fields is larger in Non-MITAP than in the MITAP papers as seen in Figure 4.6. Some of these fields are essential to report such as Storage time, Temperature, if the prewarmed medium was used and does the protocol follows Good Manufacturing Practice (GMP). On the other hand, “Individual identifier number” from MITAP top unreported

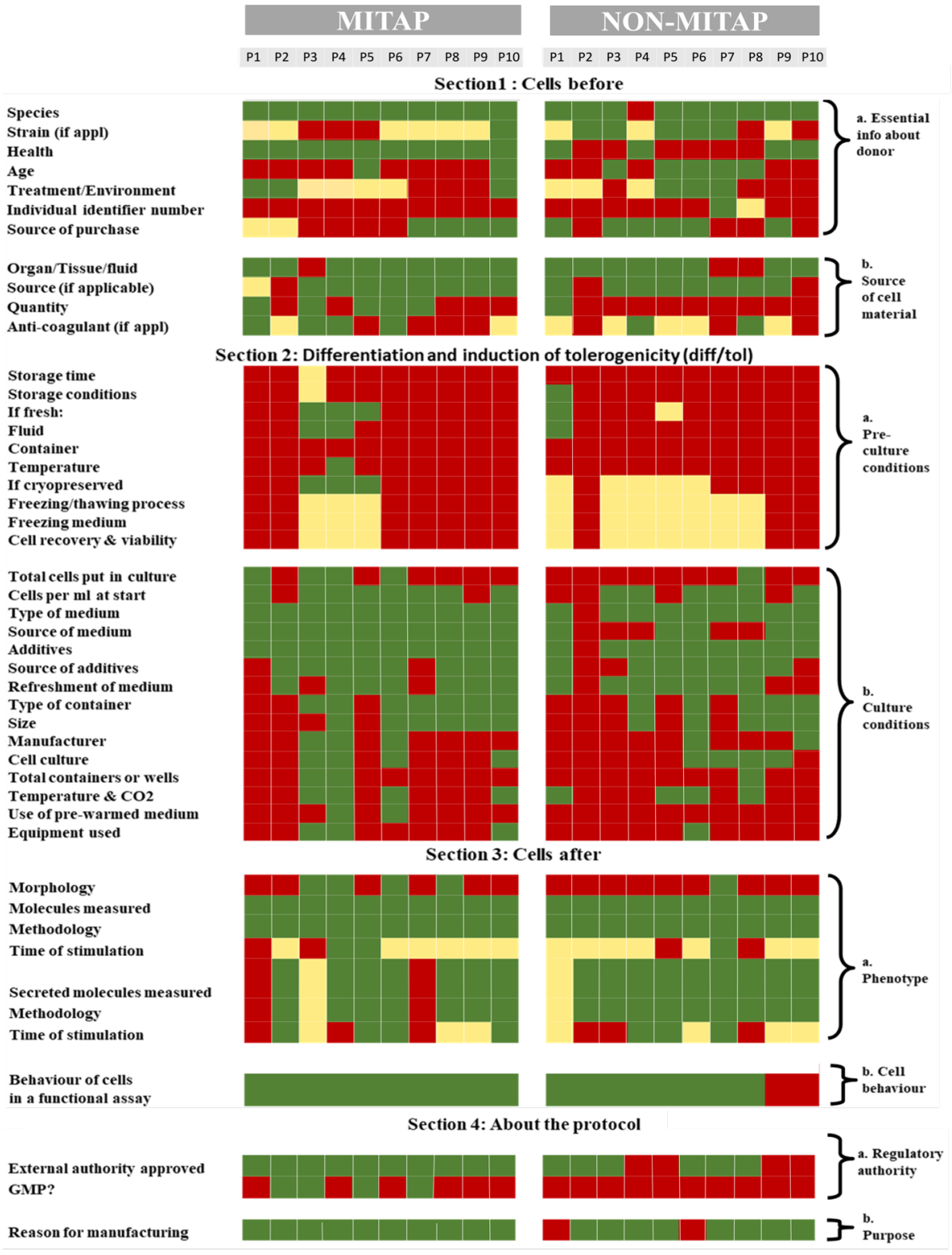
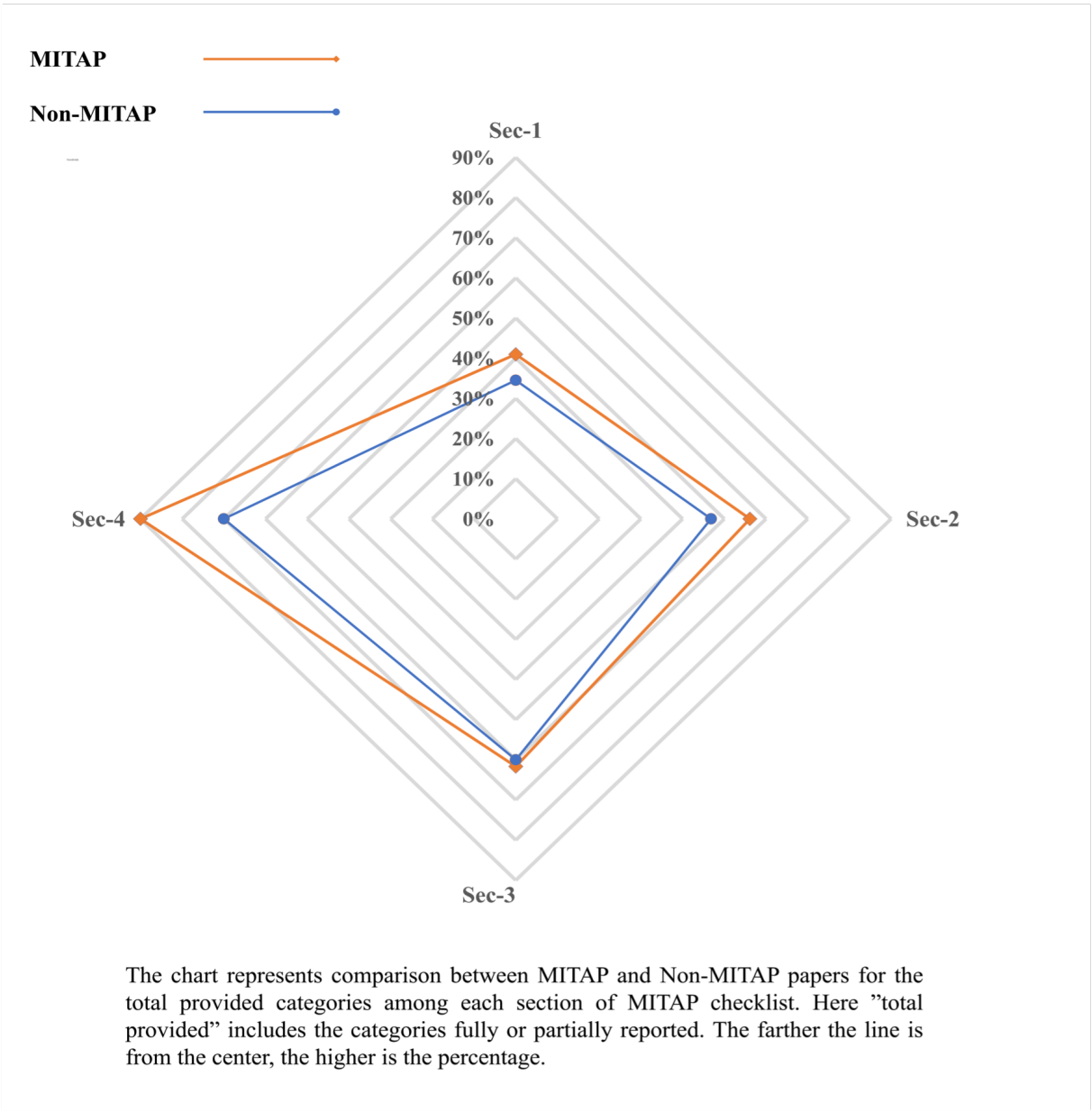


Figure 4.3: Heatplot comparing MITAP and Non-MITAP papers.
Graph shows the results of a total of 20 tolDCs papers. Green: category reported in the publication; Yellow: category partially reported in the publication; Red circle: category unreported in the publication. For the sake of clear representation, not all subsections are presented in this graph. A complete graph is provided in Appendix.



The chart represents comparison between MITAP and Non-MITAP papers for the total provided categories among each section of MITAP checklist. Here "total provided" includes the categories fully or partially reported. The farther the line is from the center, the higher is the percentage.

Figure 4.4: Graph comparing the four sections of MITAP between papers that utilized MITAP and those that did not

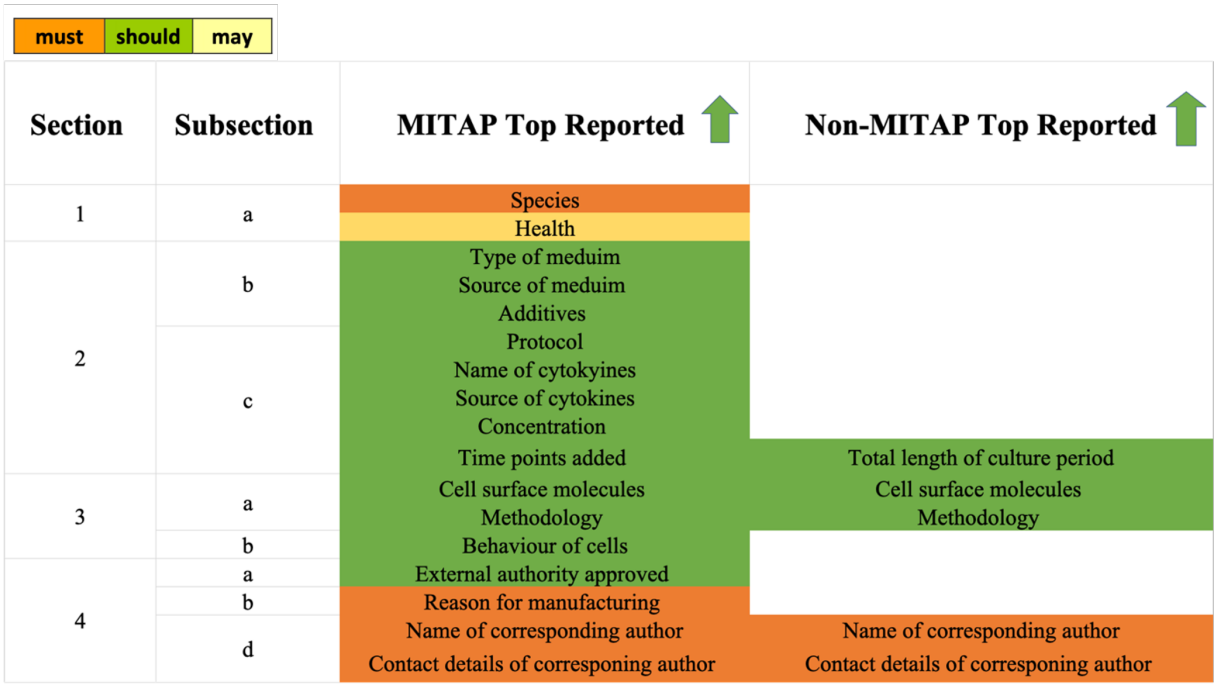


Figure 4.5: Top reported fields in MITAP vs Non MITAP.
Comparison of the most frequently reported metadata fields in papers using MITAP vs those not using MITAP. The figure highlights key areas where MITAP adoption has improved data standardisation and transparency.

is a "maybe" field.

4.6.2 McNemar test

The McNemar test is a Non-parametric test for paired nominal data. Nominal data is categorical data with 2 or more categories. The main usage of McNemar test is to calculate the change in proportion for the paired data. It has a Chi-Square distribution and is used widely in the medical field to analyse experiments.

Further analysis is carried out to check the percentage of reported parameters for each section. As the data is categorical and does not follow any distribution, the Non-parametric McNemar test was carried out, using the python 3.9 `mlxtend` library. For both MITAP and Non-MITAP papers, Section 1 is the least reported section with less than 40% of the parameters reported, whereas for all other sections >50% of the parameters were reported. MITAP papers performed marginally better than Non-MITAP papers

must should may

Section	Subsection	MITAP Top Unreported	Non- MITAP Top Unreported
1	a	Idividual identifier number	
	c		Container
2	a	Container	Storage time
			Container
			Temperature
	b		Use of pre warmed meduim
	e		Storage time
			Fuid
			Container
			Temperature
3	c		Total cell number at the end of isolation
4	a		Does protocol follow GMP

Figure 4.6: Top unreported fields in MITAP vs Non MITAP.
Analysis of the most frequently missing metadata fields in MITAP vs Non-MITAP papers. The results indicate gaps in data reporting even among studies adopting MITAP, identifying areas for further standardisation efforts.

for sections 1,2 and 4; there were no significant differences for Section 3 (Figure 4.3 and Table 4.2).

Table 4.2: Significant reporting differences between MITAP and Non-MITAP papers by McNemar's test.

Section	Subsection	
1	a	Characteristics of the organism - health
	b	Source of cell material - quantity (vol, size, weight)
	c	Cell separation - equipment used
	c	Cell separation-tissue condition
	c	Cell separation - methodology
2	b	Culture conditions - source of medium
	c	Differentiation protocol - source of cytokines/other agents
	e	Storage - Fluid
	e	Storage - Container
4	a	External authority approved
	a	Does protocol follow GMP?
	c	Allogenic/Autologous/Xenogeneic/Syngeneic

4.7 Comparison of MITAP with other MIMs

While it is known that high-quality data reporting is important and essential to reproducible research, as we have shown, MITAP is currently used by only 14% of tolDC publications. In this regard, the tolDC community does not seem usual as there has been an intense focus on the lack of reproducibility in many areas of biology and immunology. Partly this is because following standards for the design and subsequent reporting of experiments and analysis is complex. There have been noteworthy efforts made to support the data reporting. In addition, the fact that new MIMs are being introduced frequently in the medical field suggests that MIMs provide an effective way to standardise the metadata.

Because MITAP was used by a smaller number of published papers than we might have hoped, we analysed and compared the usage of other MIMs to MITAP. MITAP was published in August 2016. Figure 4.7 shows that five other MIMs were also published in 2016. Looking at the citations of these other five MIMs, MITAP performs at a similar level and is even in the top three most cited MIMs published in 2016.

In addition to comparing MITAP with MIMs in 2016, it is useful to compare it with other relevant MIMs. MIAME and MIGS are well-established MIMs in the medical field, published in 2001 and 2008, respectively [Brazma et al., 2001, Field et al., 2008]. As

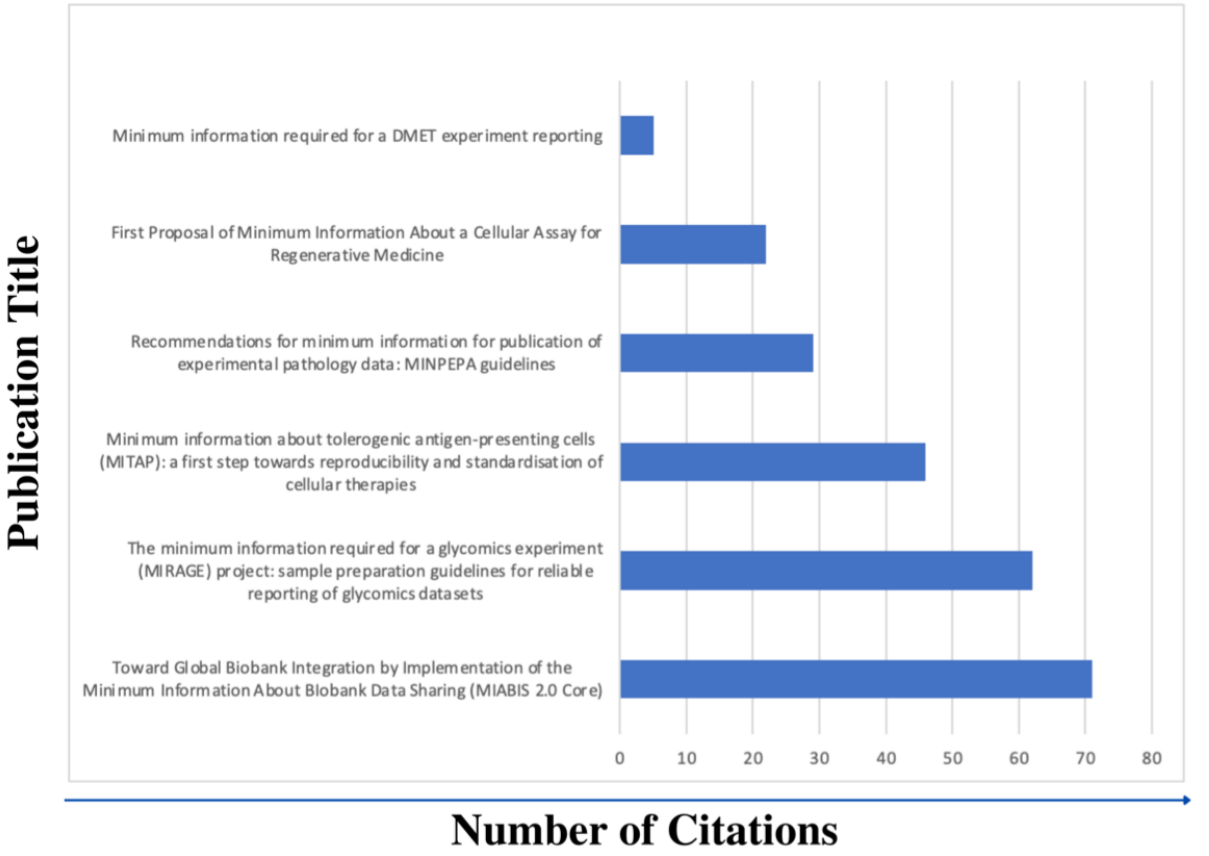


Figure 4.7: Comparison of MITAP's performance citation-wise with five other related MIMs published in the same year as MITAP which is 2016.

dendritic cells are a type of immune cells, we looked at other immunology-related MIMs such as MIATA (MIM for T cells assays) and Minimum Information about T Regulatory Cells [Janetzki et al., 2009, Fuchs et al., 2018].

Figure 4.8 shows the citation statistics of these five MIMs compared to MITAP. MIGS and MIAME are MIMs for molecular experimental data and so are rather broader, but it can be clearly seen that the citations are not huge. Similarly, MIATA and T-cells are also substantially bigger and older research fields than the tolDC field, but we see a maximum of 16 citations only in the peak year.

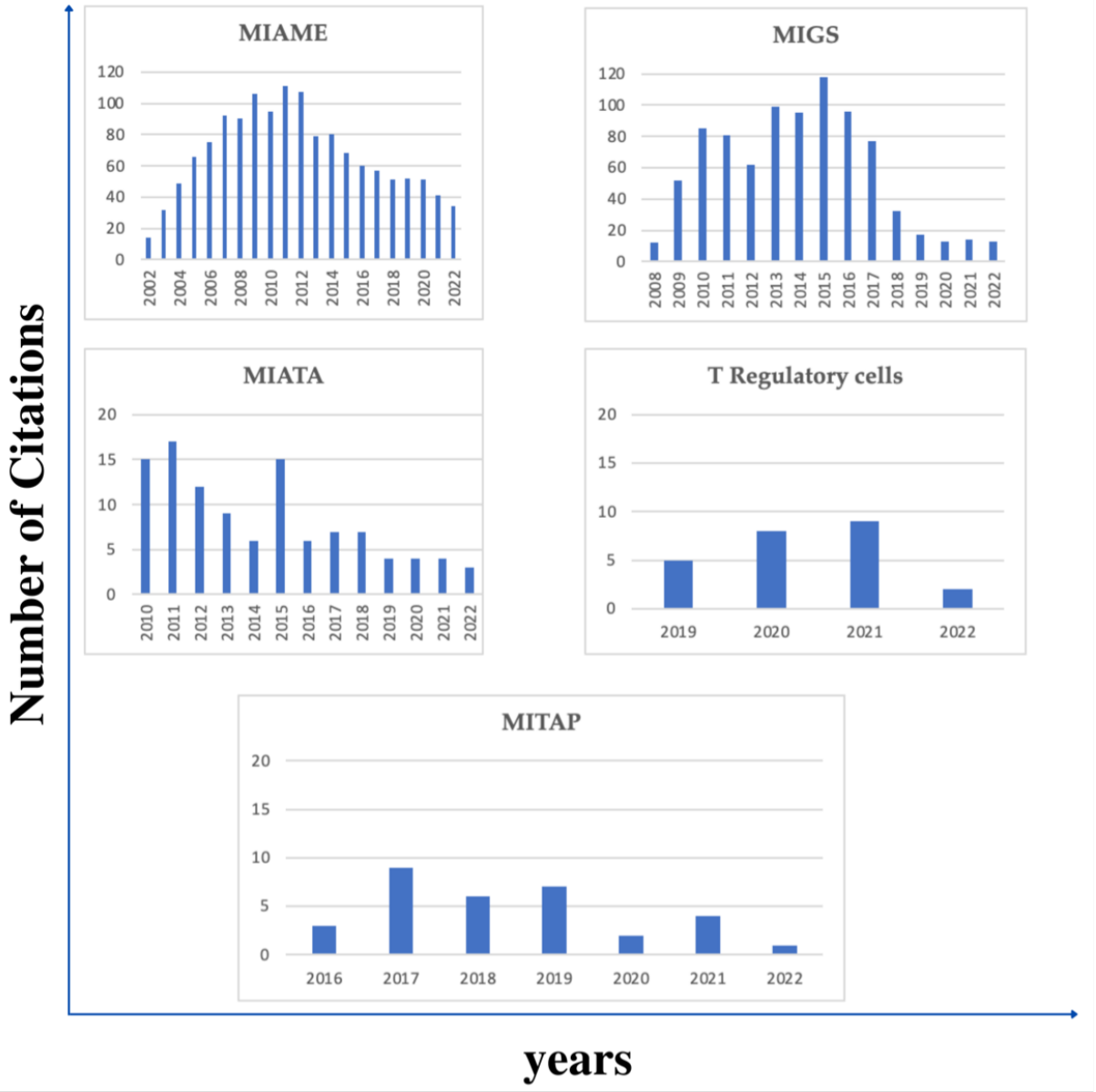


Figure 4.8: Comparison of MITAP's performance citation-wise with four other tolDC related MIMs.These MIMs are selected for the comparison because they deal with molecular experimental data or they deal with the same category of cells such as T cells.

With all these statistics, we can say that MITAP is performing well and is a sufficient tool to implement standardisation in the field of dendritic cells. Even when MITAP's checklist is not followed completely, it still fulfils the long overdue requirement of providing a minimum set of most important entities for tolDCs and increases the data reporting.

4.8 Discussion

This work shows that 5 years after its inception, MITAP is underused by investigators in the tolDC field. But where it is used, the amount of metadata available is slightly increased over where it is not. From this, we conclude that MITAP constitutes a partial success only, with further work required to improve standardised reporting in the field.

The reason for the low uptake of MITAP is unknown but may have been partly caused by using the term 'tolerogenic antigen-presenting cells' in the title. This term was deliberately chosen to cast the net wider, allowing for the incorporation of not only tolDC but also of other relevant cell types, for example, MRegs. Both tolDCs and MRegs have been developed for the induction of immune tolerance and have undergone testing in clinical trials. However, although both tolDC and MRegs are well-established and recognisable names in the community, the term tolAPC is not.

Other reasons may be that investigators perceive the process of applying MITAP to their papers as too cumbersome or unnecessary. However, MITAP does provide an easy-to-use checklist to ensure that all relevant elements of the cell production protocol are described, and our analysis of the literature show that standardised reporting is necessary.

There is a growing recognition that open data reporting and standardisation are necessary for repurposing or reusing data to discover valuable insights from past work, thus promoting research transparency. Our experience shows that the take up of MIMs takes time, despite the general appreciation of the importance of research transparency. A possible reason could be, that for the person producing the data, they are time-consuming to use, but conversely, for the people consuming the data, they provide the benefits of interoperability and reusing with minimal effort. Experts develop MIMs in the field with a lot of hard work. The primary reason that a MIM would be successful is the availability of a relevant data repository. Such as MIAME is one of the most successful MIM and has relevant repositories such as GEO to enable MIAME complaint data submission. As a result, many journals also require researchers to follow MIAME to submit the data. Other

factors to the success of a MIM can be the availability of software to record the metadata combined with a database to retrieve the data along with the metadata efficiently. Our experiment proves that MITAP is sufficient for the reporting of data. To make it more successful and accessible in the field, we would need to introduce an online platform where researchers can record, edit and save the metadata in a user-friendly manner.

Additionally, the trend in MIM adoption suggests that they require several years to become widely accepted. Furthermore, the analysis revealed that MITAP is used more in the European region and is not followed in other parts of the world. This discovery led to further exploration of the geographical impact of the tolDC field worldwide, which is presented in detail in Chapter 7.

There are many challenges associated with the MIMs following as explained in the discussion. In addition to journals implying that the researchers should follow MIMs, the researchers can be educated about the usefulness of following MIMs which is believed to be delivered through this study. Another factor in the lower following of MITAP specifically is the unavailability of a specific data warehouse. It was observed that MIMs which have specific databases for their research fields tend to be followed more than others. The next chapter 5 of this thesis is focused on the integration of the tolDC data.

4.9 Summary

Over 5 years ago, investigators in the field of tolDC came together to create MITAP, a minimum guidelines tool for reporting the protocols to generate these cells. A survey was conducted on extant papers that were published before MITAP came into existence, which showed that a large proportion of papers lacked sufficient data required to interpret and reproduce the generation of these tolerogenic dendritic cells [Lord et al., 2016]. Here, we investigated the usage of MITAP and whether it improved the description of the four protocol sections: 1) Cells before; 2) Differentiation and induction of tolerogenicity; 3) Cells after and 4) About the protocol.

Despite the highly collaborative nature of building MITAP, which involved many experts in the field, we have found that the use of MITAP is surprisingly low: only 14% of research papers that could have used MITAP, actually did. Although encouragingly, those papers that used MITAP performed slightly better on reporting relevant data across three of the four protocol sections, experimental details remained under-reported, especially in

sections 1 and 2.

There were no significant differences in reporting on section 3 ('Cells after') between MITAP- and Non-MITAP papers. This can most likely be explained by the fact that section 3 deals with the final tolDC product itself and as these cells are the main focus of these papers, details on these cells are usually well-reported. In contrast, sections 1 and 2, which describe the source and features of the cells before (section 1) they undergo the differentiation process to become tolDCs (section 2), were significantly under-reported in Non-MITAP papers compared to MITAP papers. Under-reporting of these sections will ultimately make it more difficult for investigators to reproduce published data and/or to make comparisons between different types of tolDCs. As it is becoming increasingly clear that considerable heterogeneity exists between different tolDC products, but also between tolDC products derived from different patients groups [Navarro-Barriuso et al., 2018], it is becoming even more pertinent to improve reporting on the cell source and culture protocol to generate tolDCs.

5

tolDC data integration using knowledge graph

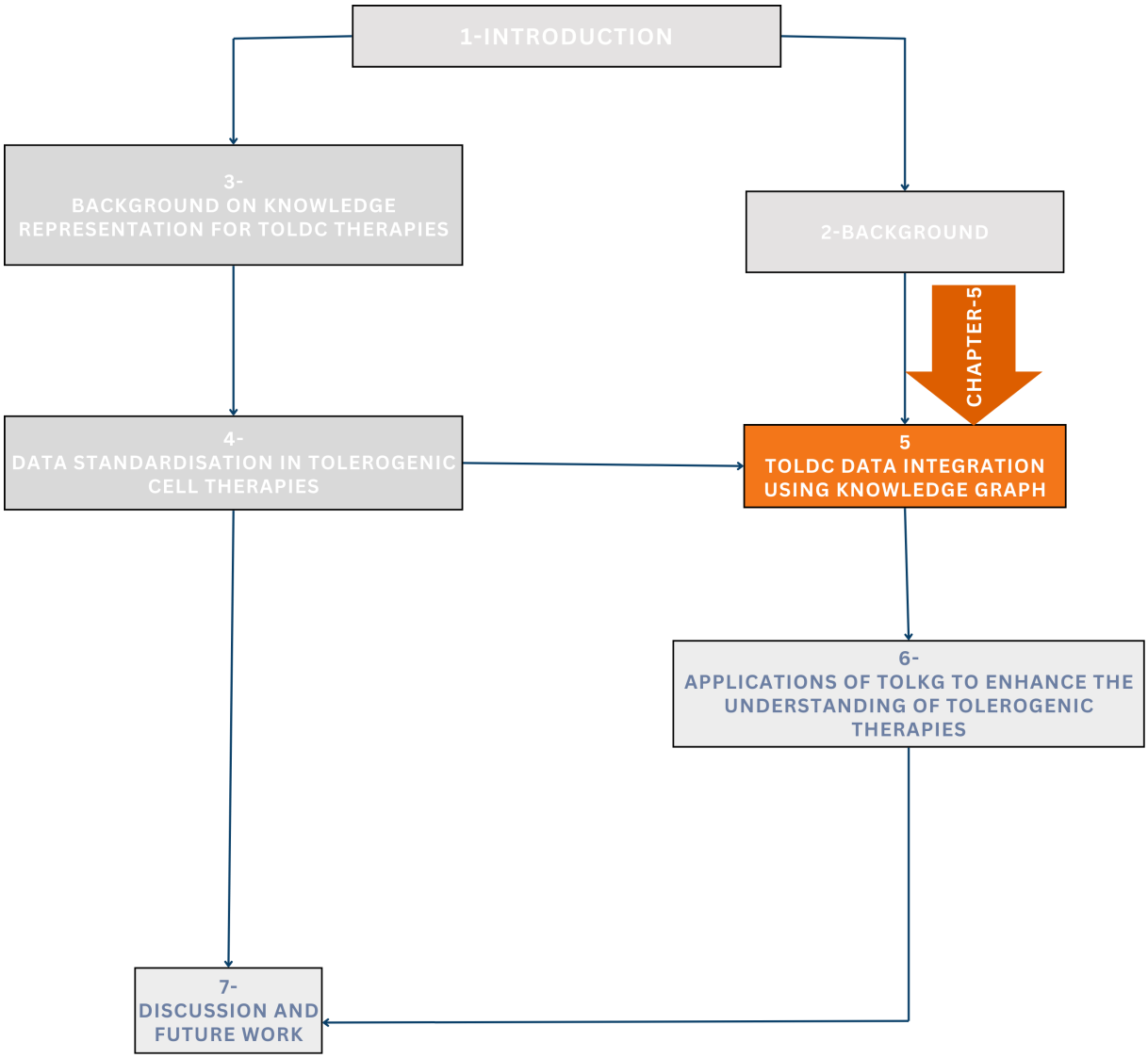


FIGURE 5.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

5.1 Abstract

We acknowledge the limited sharing of data in the field of tolDC therapies, as well as the lack of standardisation in the existing data, as discussed in Chapter 4. In this chapter, our main objective is to gather and incorporate relevant information related to tolDC therapies that are publicly available online. Our focus lies in utilising published literature to accumulate a substantial amount of background information. Our approach to data integration encompasses both structured and unstructured data sources. Instead of traditional data integration methods, we employ a knowledge graph-based approach, which allows us to explore connections and relationships within the field of tolDC therapies. As a result, we have a comprehensive knowledge graph with targeted data on tolDC therapies. This knowledge graph allows the storage and retrieval of relevant information about tolDC therapies; but more importantly, it can serve as a foundation for the generation of further hypotheses.

5.2 Introduction

One of the characteristics of research into tolDC therapies is that it is relatively new and so has fewer publications than more mature fields (see Chapter 2 for details). One of the conclusions of Chapter 4 is that we should make studies more comparable which would allow data to be re-purposed and re-used. The efficient integration of the relevant data will support this re-use. Data integration has proven to be useful in other fields. However, due to the unique challenge of the tolerogenic field, we need to adapt these approaches to address its needs.

Here, we construct a tolDC knowledge graph (tolKG) which is built upon literature and also the relevant biomedical databases. Literature was analysed and incorporated using NLP technologies over full-text articles. Other information is spread over multiple sources and needs to be integrated to capture the relationships between entities while maintaining high-quality information standards. We achieve this using a knowledge graph driven by a graph database, providing strong querying capabilities. This knowledge graph allows visualisation and analysis of different aspects of the tolDC field. The key aspects of this approach are, therefore, that this graph combines both structured data and knowledge gained from a literature corpus and it is the first study to extract and build a complete

knowledge graph in the tolerogenic therapeutics field.

5.3 Reasoning for tool selection

Before describing the method, we first explain our reasons for choosing the tools we will use and the background work involved. For entity extraction from the corpus, we are using a tool called TERMite. Prior to selecting this tool, we experimented with BioBERT, a pre-trained language representation model specifically designed for biomedical text mining. BioBERT has several advantages, such as its ability to understand complex biomedical terminologies and its robust performance in extracting entities from unstructured text [Lee et al., 2020]. However, BioBERT also has certain limitations. It requires significant computational resources for fine-tuning and may produce noisy outputs when dealing with highly specialised or less common terms that are not well-represented in its training data. Given the specificity and uniqueness of the tolDC field, along with our relatively small dataset of research papers and limited computational resources, BioBERT was not the optimal choice for us.

We also investigated lighter models such as ALBERT and SciBERT [Naseem et al., 2022, Beltagy et al., 2019]. Although these models use fewer computational resources, they also come with their own limitations, such as reduced accuracy and performance in handling domain-specific terminologies.

Based on the above discussion, we decided to use TERMite. One of the main goals of this thesis is to repurpose and reuse already available resources. Therefore, the choice of using TERMite aligns with our thesis objectives, as it leverages existing tools for biomedical entity extraction. Since developing a new tool was not the objective of this work, using an established tool like TERMite was the most suitable approach.

Evaluation of the SciBite TERMite Tool: To ensure the accuracy and reliability of entity extraction in tolKG, we conducted a systematic evaluation of the SciBite TERMite tool. Given the complexity of biomedical terminology, particularly in the emerging field of tolDC therapies, it was essential to assess the tool's performance in correctly identifying relevant entities such as genes, diseases, drugs and tolDC-specific terms.

The evaluation process involved a two-step approach. First, we performed a manual review of a subset of extracted entities, comparing TERMite's outputs against the original

text to identify potential false positives and false negatives. This helped determine the tool's ability to correctly detect and classify key terms. Second, we consulted domain experts to validate the accuracy and relevance of the extracted terms. The experts reviewed TERMite's results, providing feedback on missing or misclassified entities, particularly in specialised areas such as tolerising agents and maturation stimuli.

Overall, TERMite demonstrated strong performance in extracting standard biomedical entities, benefiting from its extensive curated vocabularies. However, some tolDC-specific terms were not initially recognised. To address this, we incorporated custom dictionaries into TERMite, improving its ability to detect specialised terminology. This evaluation confirmed that while TERMite is a robust tool for biomedical text mining, domain-specific adaptations are necessary to optimise its effectiveness for niche research areas like tolDC therapies.

5.4 Relevant data sources

In this section, we discuss the data sources to incorporate into the tolKG along with the research papers. Research papers on tolDC therapies typically mention a variety of entities related to the immune system, therapeutic mechanisms and specific biomedical elements. Genes and pathways are frequently mentioned in the tolDC research field. Therefore, we consider integrating as much information as possible about genes and pathways in the tolKG. Instead of extracting interactions among entities ourselves, we focus on utilising existing resources to incorporate these interactions into the tolKG.

We carefully selected these data sources to be incorporated in the tolKG as they are freely accessible, standardised and updated regularly.

Gene-Gene associations: IntAct is a database in the life sciences that is dedicated to the storage and management of data on protein-protein interactions. The database contains information on over 4 million protein-protein interactions and is widely used by researchers in the field of proteomics as a freely available resource for studying the functions and interactions of proteins [Orchard et al., 2014].

Gene-Drug associations: DGidb is focused on providing information on the interactions between genes and drugs and is particularly useful for researchers working in the

field of pharmacogenomics. It is a comprehensive resource that contains information on thousands of genes and drugs and includes data on the mechanisms of action, clinical relevance and potential side effects of these interactions [Cotto et al., 2018].

Gene-Disease associations: DisGeNET is a database that is widely used in the life sciences for the storage and management of information on genetic variants and their association with diseases. The database includes information on over 20 million genetic variants and their association with over 55,000 diseases [Piñero et al., 2016].

Pathways associations: Reactome is a comprehensive database of biological pathways and processes in the life sciences [Jassal et al., 2020]. It is a widely used resource for researchers in the field, as it provides detailed information on the molecular mechanisms underlying various biological processes. The Reactome database is organised as a network of pathways, with each pathway representing a specific biological process or series of events. The pathways are annotated with detailed information on the molecules and reactions involved, as well as their relationships and interactions.

5.5 Method

In this section, we aim to describe the rationale for our approach in integrating tolDC data. We recognised that the availability of structured data for emerging fields, such as tolDC therapies, may be limited or not very specific. Therefore, we turned to the scientific literature to identify relevant information that could be used to filter the structured data. By using this approach, we aimed to integrate all the available data in a comprehensive and targeted manner, enabling us to build a more accurate, informative and useful knowledge graph. This strategy allowed us to ensure that the data used for our analysis was highly relevant, reducing the potential for any extraneous or irrelevant information that may have been included had we used a broader approach to data integration; as a useful side-effect, it also reduces the computational requirements for using this knowledge graph.

The two main steps of the tolKG construction are integrating unstructured data from full-text articles and structured data from other databases as shown in Figure 5.2. While TERMite provides a foundational platform for extracting relevant entities from research

articles, the process requires additional work to ensure accuracy, relevance and meaningful integration into the research which is explained in Section 5.5.1.

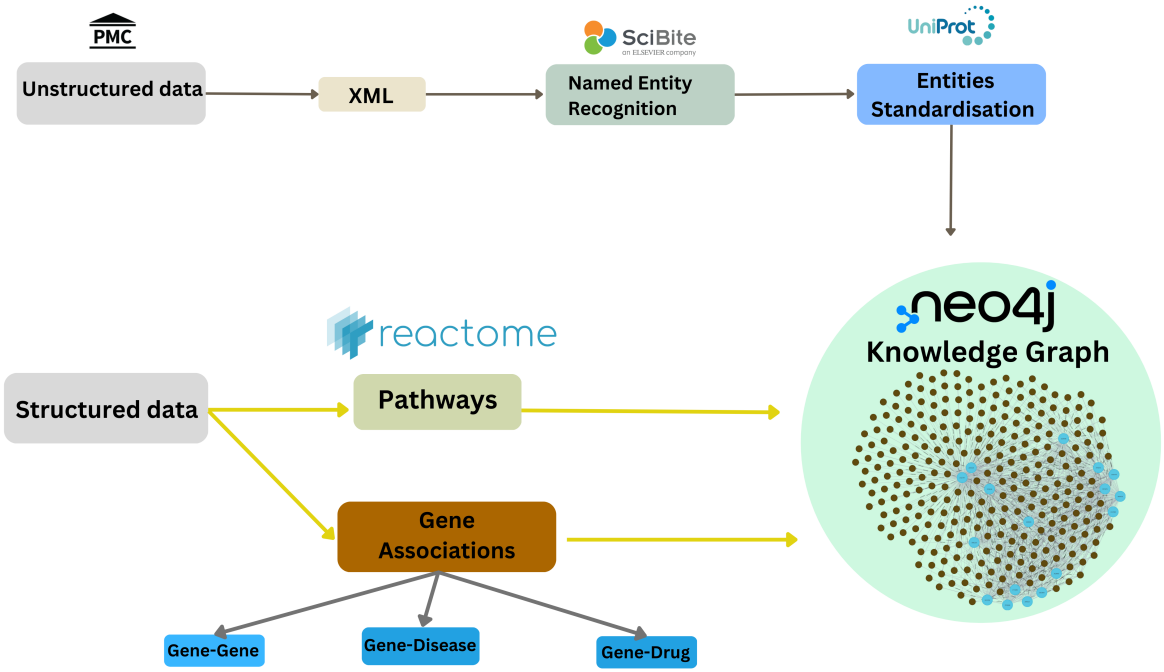


Figure 5.2: Framework for tolKG construction
Integration of knowledge from unstructured text and structured relevant biomedical databases

5.5.1 Biomedical knowledge retrieval and extraction from literature

Refining the corpus: To ensure that our data extraction and integration strategy was focused on the most relevant information, we began by refining the corpus of research articles on tolDC. This involved a careful selection process that identified and included only articles that were directly relevant to the field of tolDC therapies. To search for the publications on tolDC, we used PubMed Central (PMC) which is a free online resource providing access to free full texts of biomedical published articles [Sayers et al., 2010]. Many free full-text publications on tolDC can be found on PMC; this makes it suitable for creating the tolDC-relevant corpus.

The use of MeSH (Medical Subject Headings) in literature searches has become increasingly common in the biomedical research community [Gan et al., 2019, Riccaboni and Verginer, 2022]. MeSH is a standardised vocabulary developed by the National Library of Medicine (NLM) to aid indexers in accurately describing the content of articles in a consistent manner [Darmoni et al., 2012]. By applying MeSH terms to articles, indexers create a more precise and comprehensive database of literature, which facilitates more accurate literature searching and retrieval. MeSH terms are hierarchically organised, with broader terms encompassing more specific ones and subheadings can be used to further refine the search. This hierarchical organisation allows for greater precision in literature searches, enabling researchers to find more relevant articles more quickly. Subsequently, after thoroughly examining several research papers in the field of tolDCs, we identified "Dendritic Cell\immunology" as the optimal MeSH term to encompass the subject area of tolDCs.

To download all the research papers against this MeSH term, we developed a Python application to access the full texts of articles via the RESTful API of the PubMed Central database. To refine our search, we specifically targeted publications from the year 2000 up to the present date of February 2022. A total of 7343 full-text articles in the Extensible Markup Language (XML) format were successfully downloaded and incorporated into our subsequent analyses.

Pre-processing the corpus: After collecting all the tolDC papers, some preprocessing steps were performed before using the XML files for data extraction.

- The XML files, which contain the full text of research papers including references, can have multiple key terms in the titles of publications in the reference list. These terms or relationships can eventually become associated with the publication that cited them. Therefore, all references were excluded from the corpus as a data filtering step.
- Resolve acronyms (e.g. tolDC to tolerogenic dendritic cell). This was done using our simple acronym resolver tool based on the schwartz hearst algorithm [Schwartz and Hearst, 2002].
- Use lowercase for all text to avoid having different terms for the same word with different cases (e.g. "Brain" and "brain"). This might cause a problem if applied before resolving acronyms.

Text mining on the refined corpus: After refining and pre-processing the corpus, the next step was to extract useful entities from the text. Biomedical research papers are complex documents that contain a vast amount of information related to various biological processes and systems. Another issue which is more prominent in the biomedical field than in other research areas is the frequent introduction of new terminologies or the renaming of existing terminologies, particularly in the case of drugs, genes and viruses. In order to extract and analyse the required information efficiently from biomedical publications, natural language processing (NLP) techniques such as named entity recognition (NER) have been widely used [Nadif and Role, 2021]. However, due to the complexity and heterogeneity of biomedical texts, no single NER technique has been found to be universally effective. Instead, different NER methods and tools are typically used for different types of biomedical documents and domains [Goulart et al., 2011]. This can include specialised algorithms that are tailored to specific types of entities or relationships [Popovski et al., 2019], as well as machine learning approaches that can learn to recognise patterns and structures within the data [Peng et al., 2020, Song et al., 2021]. Despite these efforts, the development of accurate and reliable NER tools for biomedical research papers remains an active area of research, with ongoing efforts to improve their performance and adaptability to different types of data and research domains.

In this study, we took advantage of a tool called TERMite, developed by SciBite in Cambridge, UK, for the purpose of facilitating analysis in the biomedical domain.

TERMite is built on manually curated vocabularies that encompass over 20 million synonyms and it has been successfully applied in diverse applications such as drug repurposing [SciBite, 2023]. Further details regarding the use of the tool can be accessed in related white papers available on the SciBite website. In this study, TERMite performed adequately for the intended purpose and helped to streamline the analytical process. Additionally, the tool allowed us to integrate very specific vocabularies into the tolDC field as well.

TERMite extracted the most important entities such as genes, diseases, drugs, proteins, chemicals and lab procedures. However, some entities specific to the tolDC field were not tagged, namely cell markers, tolerising agents and maturation stimulus. These were tagged using dictionary-based NER and are available at the GitHub repository. The quality of the extraction process was checked by manual inspection of a subset of the extract entities. Furthermore, the correctness of the entities can also be proven by running different queries; examples are shown in Chapter 6. The statistics of these entities are shown in Figure 6.6.

Standardisation of entities: Entities within a text or dataset can often be complex and multifaceted, leading to potential ambiguities and inconsistencies in reference and interpretation. This complexity arises from two key linguistic phenomena. First, there is the issue of *synonyms*, where a single entity may be referred to by several different terms that have the same or similar meanings. This can create confusion in identifying and tracking the entity across different contexts or texts. Second, there is the challenge of *polysemy*, where a single term can have multiple meanings or refer to different types of entities depending on the context in which it is used. This can lead to misunderstandings and misinterpretations of the intended reference. Given these challenges, it becomes essential to implement a normalization process for the extracted entities. This process aims to standardise the representation of entities, resolving synonymous terms to a common reference and clarifying the intended meaning of polysemous terms. By doing so, the normalization ensures a consistent and unambiguous understanding of the entities, facilitating more accurate analysis, interpretation and utilization of the information.

Different normalisation methods have been developed by the researchers to resolve the ambiguity issues in the biomedical literature mining such as GNorm plus for the genes/proteins [Wei et al., 2015]. Here we utilised multiple normalisation methods to

standardise the extracted terms. The details are shown in Table 5.1.

Table 5.1: Normalisation methods for the standardisation of entities

Entity type	**Model**	**Dictionary**
Genes/ Proteins	GNormPlus	Entrez Gene
Disease	Sieve-based entity linking	OMIM
Drug/ Chemical	tmChem	DrugBank, ChEBI

5.5.2 Structured data integration

After extracting tolDC-specific knowledge from unstructured text, large-scale structured knowledge from existing biomedical resources was also integrated into the knowledge graph which provides the background structure for the database and facilitates integrative in-depth analysis using tolKG. Specifically, we have added the following four types of resources.

- **Pathways** were incorporated from Reactome to build the relationships between genes [Fabregat et al., 2018]. Firstly, a subgraph of Reactome Neo4j graph is built which consists of all pathways that contain at least one gene found in the unstructured corpus. Then, we imported this subgraph into the tolKG.

- **Gene-gene** associations were found by batch searching the IntAct API [Kerrien et al., 2012]. Firstly, a gene list was established by removing all duplicate genes. This gene list was passed to IntAct and all interactions where the MI (Mutual Information) score is greater than 0.5 are accessed. The MI score of greater than 0.5 was chosen to integrate only highly confident interactions. An edge was created if interacting genes appeared in the knowledge graph.

- **Gene-disease** associations were incorporated by pragmatically accessing the DisGeNET API [Piñero et al., 2016]. The UniProt IDs of genes that appear in the NER are passed to the DisGenNet API that returns the Gene associations, which were used to create edges in the knowledge graph where relevant. If the disease ID against the gene ID is not found in the extracted entities from the literature, the disease associations are added by creating additional disease nodes.

- Lastly, **Gene-Drug** interactions were accessed via the DGIdb (Drug Gene Interaction database) API [Cotto et al., 2018]. The CHEMBL IDs of extracted drugs, after removing duplicates, are passed to the API which returns the gene-drug associations. Similar to the gene-disease association if the associated gene is not found it is still added as an additional gene node in the knowledge graph.

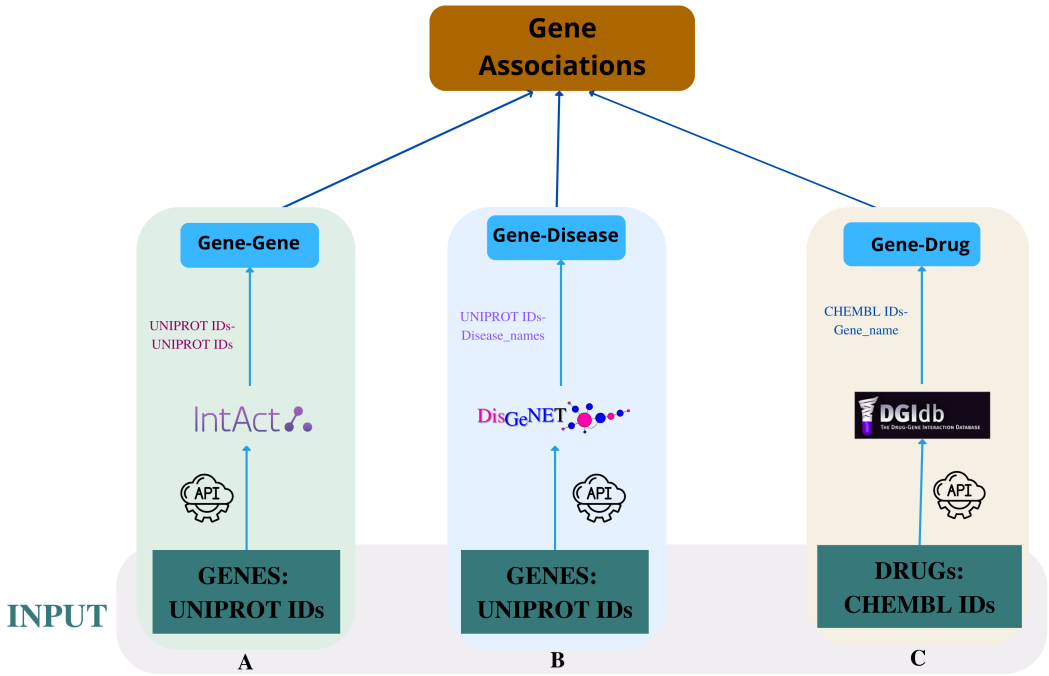


Figure 5.3: Layout of structured data integration into tolKG by using APIs of IntAct, DisGeNET and DGidb.

5.5.3 Research Paper Metadata Extraction and Classification

Author information extraction from the corpus: Extracting author information from a corpus of XML files is a multifaceted task that begins with understanding the XML schema, which defines the structure of the XML documents and includes tags and attributes holding relevant information. The process then involves parsing the XML files using specialised libraries available in programming languages such as Python, Java, or C. After parsing, the next step is to navigate to the specific tags containing author information, such as names, affiliations and emails and extract these details. When dealing with a collection of XML files, a loop or batch process must be implemented to iterate through each file and apply the extraction process uniformly. The extracted information may then be stored or outputted in various formats like CSV, JSON, or a relational database, depending on the requirements. It is also crucial to consider any legal or ethical guidelines, especially when handling personal or sensitive data. For example, using lxml library in Python, one could parse an XML file, iterate through author tags and print the name, affiliation and email. The process, although systematic, requires careful attention to the specific structure of the XML documents and the choice of tools to ensure efficient retrieval and utilization of the author's information from the corpus of files.

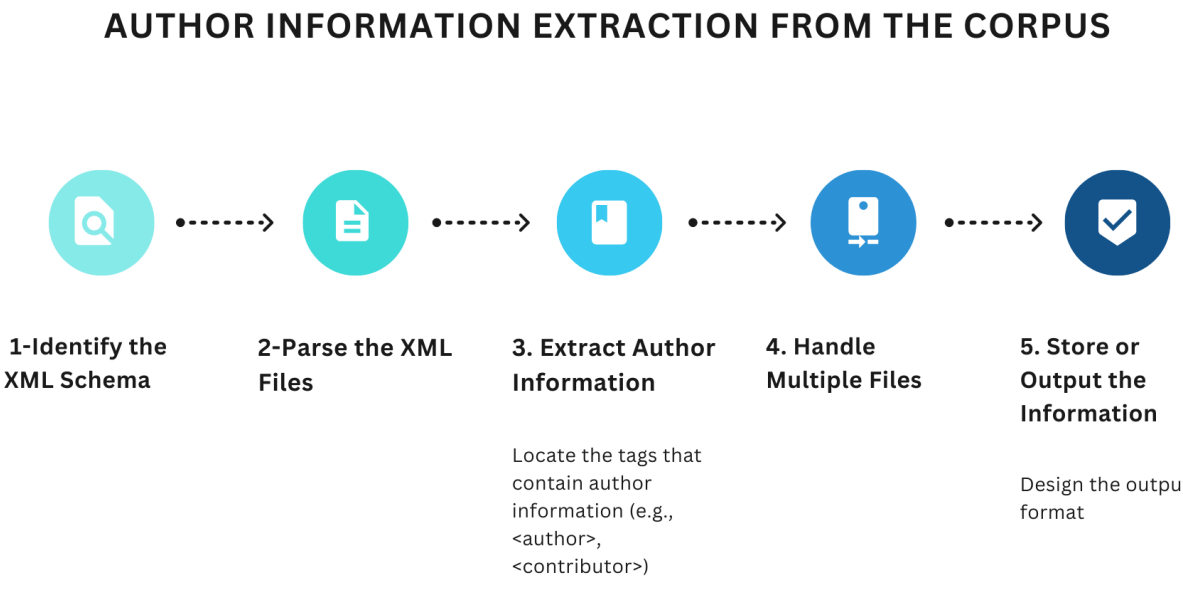


Figure 5.4: Framework for author information extraction from XML files

Research paper classification: We have also categorised the research papers into publication types such as review articles, methodology papers or clinical trials etc. The categorization of research papers into these specific types holds significant importance for various reasons:

- **Facilitated Comparison:** By grouping papers according to their type, researchers can more easily compare papers within the same category. For example, if researchers are interested in comparing the methodologies used across various studies, they can focus specifically on the set of methodology papers.
- **Tailored Analysis:** Different publication types require different approaches for reading and analysis. A review article, which summarizes existing research on a topic, might be approached differently from a clinical trial report, which presents original experimental data. By categorizing papers, we can apply the most appropriate analytical tools and frameworks to each type.
- **Improved Search and Retrieval:** The categorization into publication types also significantly enhances the efficiency of search and retrieval processes. Finding the right type of paper becomes a more streamlined task.

The pipeline used to perform this categorization often resembles the one employed for extracting author information from XML tags. It involves parsing the XML files, extracting information related to publication type, classifying the papers into corresponding categories and storing or outputting the categorized data for further use. Using the strategy shown in Figure 5.4, we could extract the general categories of papers, which are review articles, research articles, reports, letters, brief reports, discussion papers, editorial papers, article commentaries, case reports and clinical trials.

For biological fields like tolDC, many papers focus on the protocols to generate these cells. If such protocol papers can be classified separately, it could be really useful for the analysis and comparison. Thus, we further categorise the research article type, which is basically original papers, into *"protocol paper"* category.

We use doc2vec to categorise the protocol papers [Le and Mikolov, 2014]. While performing the analysis on MITAP, shown in Chapter 4, we manually selected the papers that are protocol papers. We took advantage of those 20 protocol papers to find similar papers to them. Because we are dealing with a specific field, we chose doc2vec for

generating document-level embeddings to train from scratch.

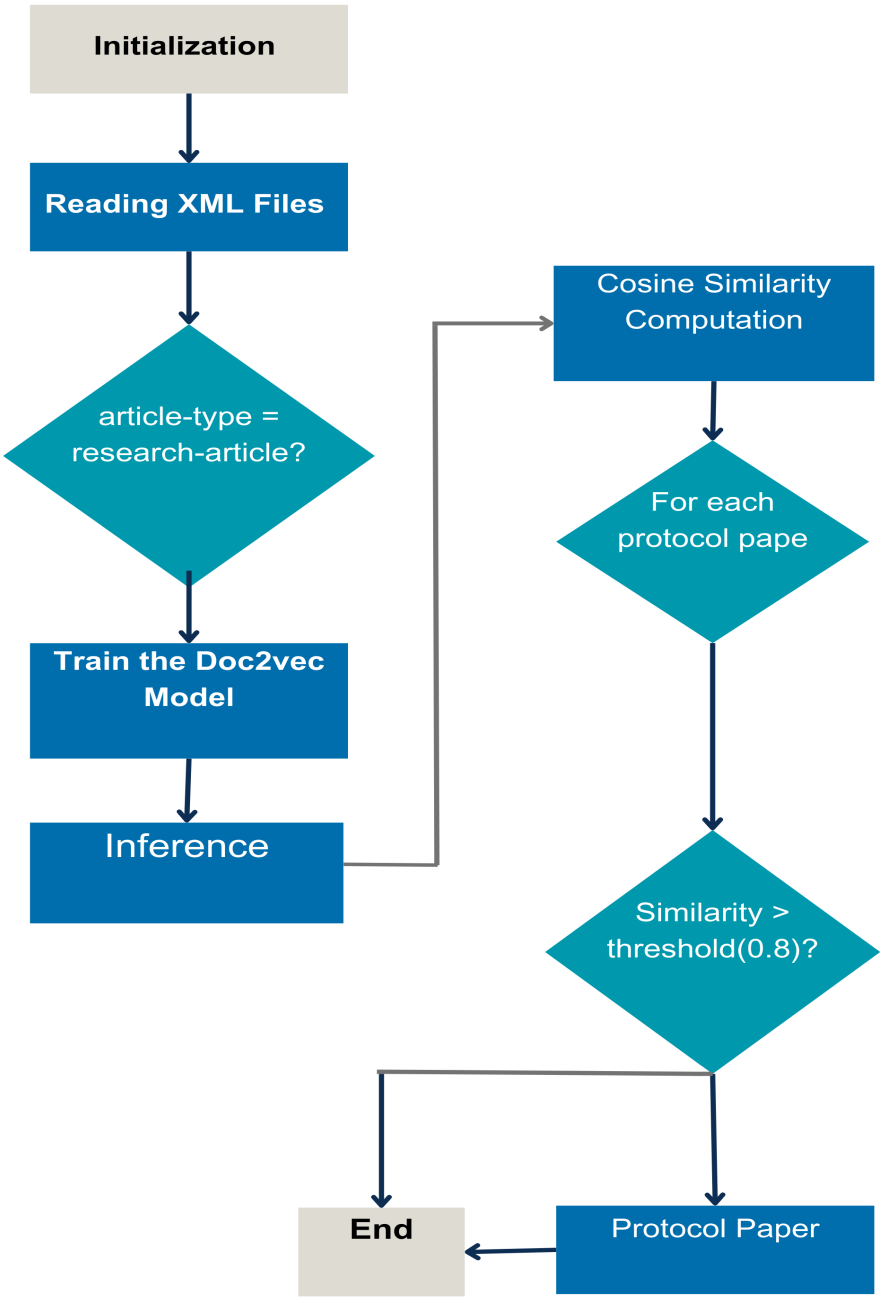


Figure 5.5: Pipeline of research paper categorisation.
This criteria of 0.8 is selected after going through the retrieved research papers with an expert in the field of tolDC.

Gender determination based on names: We used the Gender API to determine gender based on the names of the authors. This database holds 6,084,389 validated names from 191 countries and 6,196,452 unique names globally. While the database provides an accuracy of 95%, it is not flawless. We also incorporated the country of the institute with author names to enhance precision. Nonetheless, challenges persist as some names like Andrea can signify different genders in different countries and many researchers work outside their home countries, adding complexity to gender determination. Gender API also provides an accuracy value for the gender ranging from 1 to 100. We also used the results where the accuracy value was above 80%.

5.5.4 Data modelling strategy

Once we had collected all the necessary data, our next step was to integrate it into a database in a useful manner. Our data was multi-source and semi-structured, which made the graph data model a more suitable choice than the relational Structured Query Language (SQL) data model. NoSQL databases are also known for being scalable and flexible [Wang et al., 2014]. Moreover, graph traversal-type queries are particularly useful for hypothesis generation, as they can reveal relationships connecting entities that might not be expected or apparent from a visual inspection of the network. In this study, we chose to use Neo4j [Miller, 2013] as our graph database, as it is a well-established and widely used tool for data integration in the biological domain. Neo4j stores data in the form of nodes and relationships, with each node and relationship having its own properties. This database technology is being extensively utilised in the biological domain and is particularly noteworthy for the Neo4j conversion of Reactome, which provides a graphical representation of pathways [Fabregat et al., 2018]. The Cypher queries used to achieve this are available as supplementary data.

In the context of our study on tolDC therapies, physical entities such as genes, cell markers, diseases, drugs, lab chemicals and pathways are the primary focus of our analysis. These entities can belong to multiple thematic sets and thus, can be grouped by disease involvement, tissue and metabolic pathways, or linked to particular diseases. To account for the potentially complex inter-relationships between these sets, we represent these entities as nodes in our graph data model, which can be directly referenced in query patterns. Additionally, edges in the graph are used are used to indicate both i) the

membership of an entity to a set such as in publication as well as ii) the relationships between entities such as gene-gene etc. A conceptual model of how different types of entities identified through NER in scientific papers are represented in the tolKG is shown in Figure 5.6. The model highlights the connections between a paper and the various entities it mentions, as well as the basic bibliographic information associated with the paper. An actual snippet of the toKG is represented in Figure 5.7 where a cypher query is executed to show all the research papers in the tolKG but limit the returned graph to one hundred nodes only for visualisation.

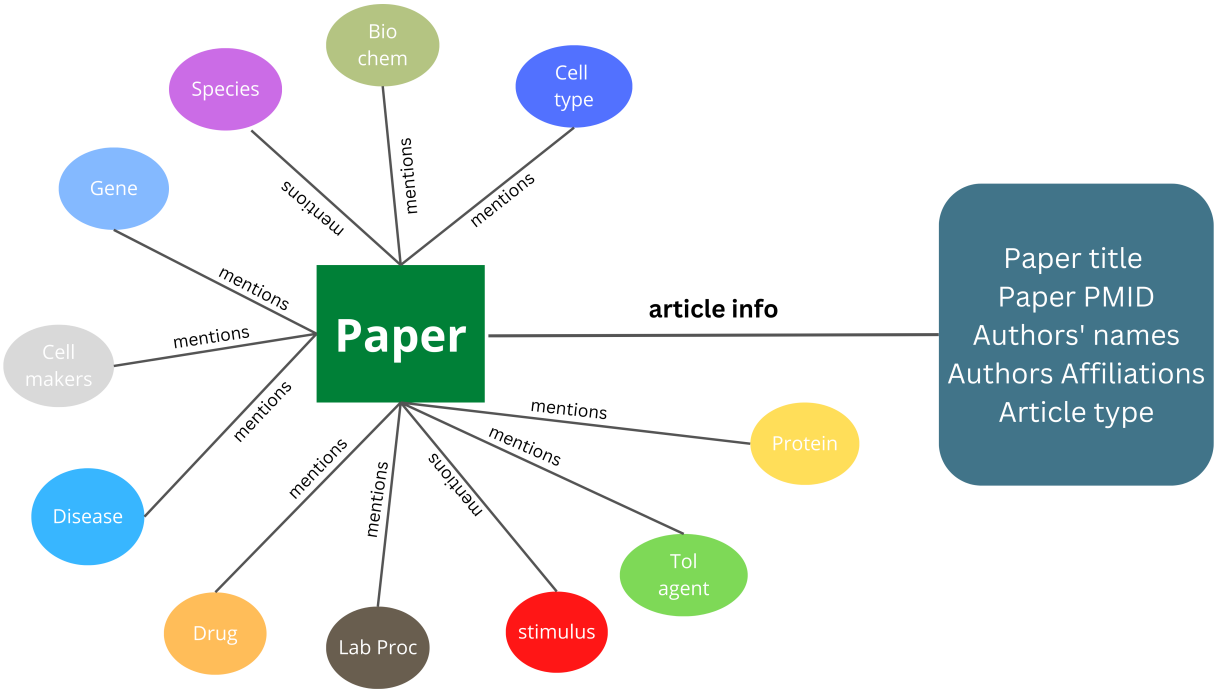


Figure 5.6: Entities and relationships in tolKG

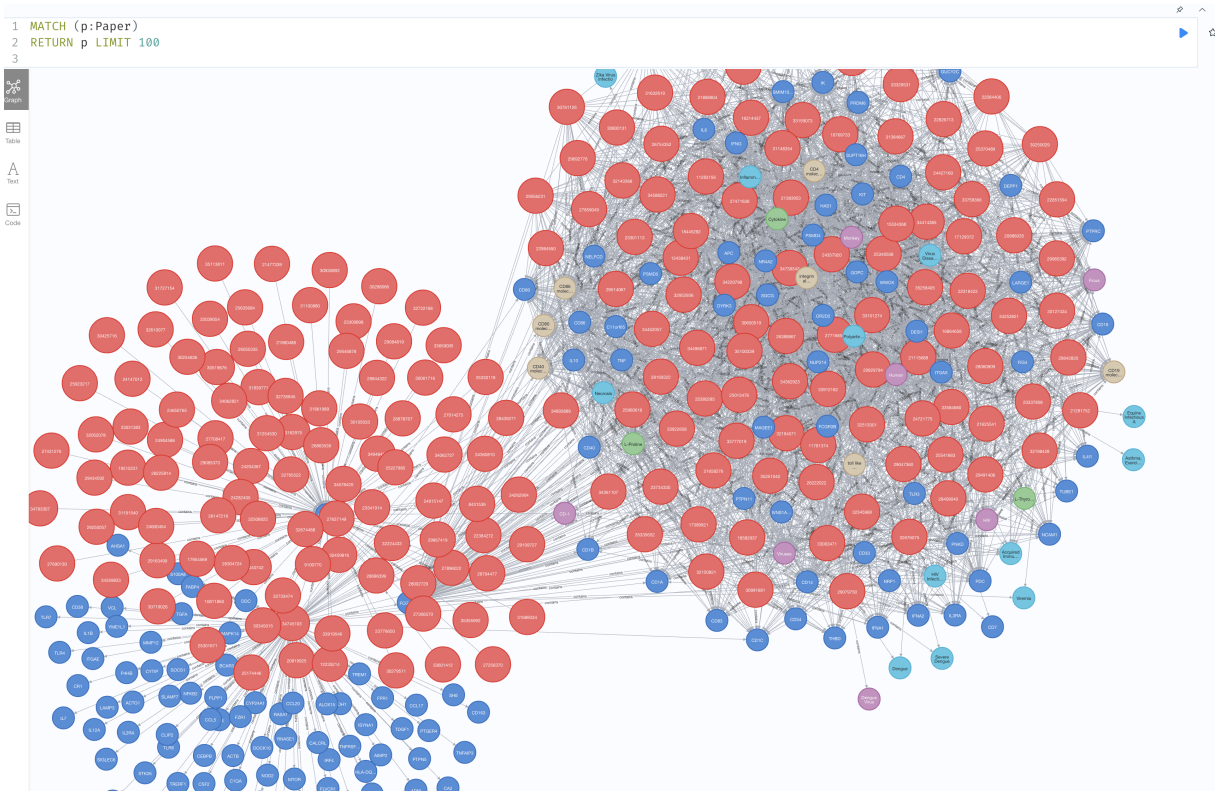


Figure 5.7: A representation of a subset of tolKG

5.6 Results

A hybrid corpus of tolDC literature was built with unstructured published data and integrated with structured data. These published articles usually contain motivation, new hypothesis testing, state-of-the-art methods, results and conclusions. TerMite tool extracted important entities mentioned in these published papers. As a result of NER, 120k entities were extracted, including genes, proteins, drugs, diseases, lab procedures, etc. Figure 5.8 provides the percentage of entities extracted from the literature.

We extracted key entities (Genes, Cell Markers, Diseases, Drugs, Protein Types), including specific tolDC entities such as tolerising agents and maturation stimulus (Maturation Stim) as well. Lab Proc are laboratory procedures such as flow cytometry, ELISA and so on. The species in Figure 5.8 represent the organism mentioned in the protocol. Cell type indicates the type of cells such as T cell, B cell etc., while Bio-Chem is used for the different chemicals mentioned in the studies, such as lipopolysaccharide.

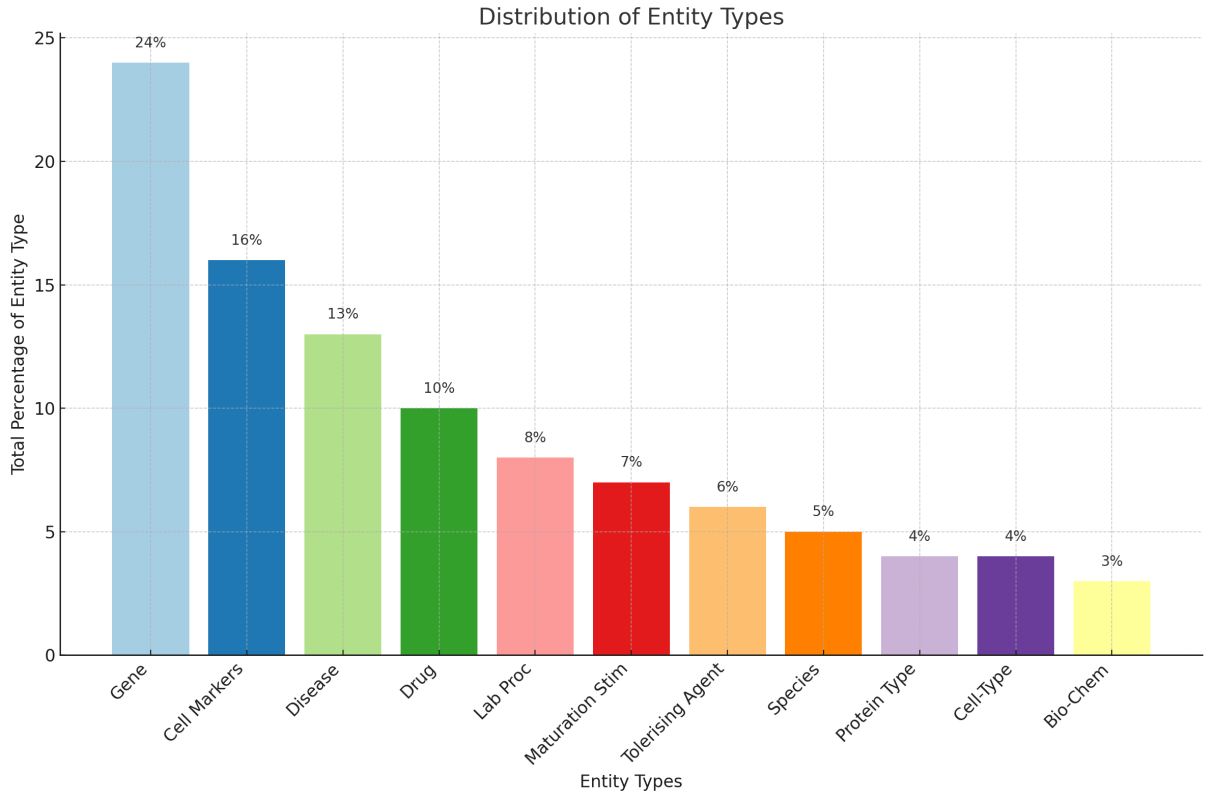


Figure 5.8: Graph showing the percentage of entities extracted from the corpus Distribution of extracted entities from the tolDC-related literature corpus. The figure shows the percentage of different entity types (e.g., genes, pathways, drugs) identified through text mining and integrated into the tolKG knowledge graph.

tolKG is a comprehensive graph and contains useful relationships between entities. The edges between the nodes represent the relationships extracted. Around 92k relationships are extracted by integrating structured resources. As a result of structured data integration, a total of 50,576 edges are created between Gene-Gene, which shows that these two genes are linked. Gene-disease and Gene-Drug have relatively smaller number of edges, which are 20k and 18k, respectively, as shown in Table 5.2.

Edge type	**Number of edges**
Gene-Gene	50k
Gene-Drug	18k
Gene-Disease	20k

Table 5.2: The table shows the relationship types between edges and the statistics about these edges.

A total of 7,343 full-text articles in the XML format were successfully downloaded and incorporated into our subsequent analyses. We categorized these papers into various research paper categories. The categorization process involved two steps: first, extracting the research paper categories mentioned in the XML tags, and second, identifying the *protocol papers* by training a doc2vec model. Table 5.3 presents the distribution of the different paper categories among the total 7,343 papers.

Paper Category	**Percentage of Papers (%)**
research article	77.0%
review	11.5%
protocol papers	8.9%
brief report	0.9%
editorial	0.9%
article commentary	0.2%
letter	0.2%
discussion	0.2%
case report	0.1%
report	0.1%

Table 5.3: Percentage of Papers by Paper Category

5.7 Dumping the tolKG data into an RDF exposed data set for public access

To improve accessibility and usability, we have converted the tolKG data into Resource Description Framework (RDF) format and made it available as a public dataset. Within Neo4j, tolKG occupies 1.6 GB on disk and takes 100 minutes to build. With the aim of offering an accessible interface to the community, an RDF rendition of the graph was generated using *NeoSemantics*. In this RDF format, the amalgamated dataset comprises more than 5.5 million triples and occupies a disk space of 360 MB.

NeoSemantics is a plugin for the Neo4j graph database that facilitates working with RDF data and concepts. RDF is a standard for representing and linking structured data on the web. The NeoSemantics plugin extends the capabilities of Neo4j by allowing users to convert Neo4j graph data into RDF format

The data can be accessed at this GitHub repository.

5.8 Discussion

5.8.1 Text mining

In this part, our focus is to retrieve important entities out of the research articles only and not the phrase detection. The entities extracted are mostly biological except for the author information. We have used a number of strategies for the retrieval of all these entities. To retrieve genes, diseases, drugs etc, we utilised the Scibite tool and after manually looking at these entities, we could see that it was able to retrieve almost all important entities mentioned from the text, which is better than training the Biobert model. The effectiveness of the Scibite tool is further enhanced by its use of manually curated dictionaries, which are both rich and up-to-date. Our objective here was clearly not to fine-tune or develop another machine learning model like Biobert specifically for this task. Using the Scibite tool provided us with a faster and richer resource to extract entities. On the other hand, the complexity is low as with the Biobert approach it requires more resources computationally and is usually slower. However, we were able to tackle the synonyms also which made it a better solution.

To extract more specific chemicals and tag them according to the specific field such as

tolDCs, we built our own dictionaries also. One of the examples is the maturation stimulus and cell markers. We created a dictionary of the most commonly used maturation stimuli along with their possible synonyms by asking within the group as well as utilising online websites that sell these chemicals. To create a dictionary for the cell markers we used the cell markers list from cellmarkersdb website.

The third approach for the non-medical entities such as the author information is the use of XML tags. In the XML tags of the author names, names are mostly only the initials. So we could extract the initials only. Another issue, we had to deal with is related to specifying the number of authors as all papers have varying numbers of authors. It is straightforward to extract the first author from the list. We developed a Python script that counts the number of items in the list and provides the exact number of authors. After assigning the author number to the authors, we could easily extract the last author.

One main point we understood from this experiment is that no single technique can serve the purpose of extracting all kinds of entities from the text. Whether dealing with complex biological data, literary analyses, or technical documents, the diversity and intricacy of language present challenges that often require a multifaceted approach. Different methods excel in different contexts. For example, machine learning models like BioBERT may be adept at understanding context and nuances, capturing entities that are not explicitly defined. In contrast, dictionary-based approaches offer precision and speed for well-defined entities but may lack flexibility. The experiment underscores the importance of leveraging a combination of techniques, possibly integrating statistical, rule-based and machine-learning methods, to create a more comprehensive and adaptive entity extraction system. This hybrid approach can cater to the varying demands of different texts and domains, enhancing the robustness and accuracy of the entity retrieval process.

5.8.2 Unstructured data integration

Many research studies are focused on extracting relationships between biological entities such as gene-gene interactions from the text. This pursuit is vital in understanding complex biological networks and functions, as it enables researchers to understand the intricate relationships between genes, proteins and other cellular components. The extraction of these relationships often involves sophisticated text mining and computational

techniques tailored to interpret scientific literature and databases. However, here we took advantage of the already extracted relationships for the tolKG. Leveraging pre-existing knowledge and previously identified relationships can significantly enhance the efficiency and accuracy of new research. We understood that by building on established connections and validated data, researchers can more rapidly explore specific subfields, such as particular disease mechanisms or targeted therapeutic interventions. This approach not only saves time and resources but also fosters a more integrated and cumulative understanding of biological systems, bridging gaps between different research areas and contributing to a more holistic view of life sciences.

5.8.3 Ensuring data accuracy and reliability in tolKG

In constructing the tolDC Knowledge Graph (tolKG), ensuring high data quality is essential. The process of extracting knowledge from literature, integrating structured data and normalising entities introduces the risk of false positives (incorrectly identified relationships or entities) and false negatives (missed relevant entities or relationships). Both types of errors can significantly impact the reliability of the knowledge graph and its downstream applications, such as hypothesis generation and comparative analysis. This section outlines the strategies used to minimise false positives and false negatives at different stages of tolKG construction, including entity recognition, standardisation, relationship extraction and knowledge integration.

To ensure the accuracy of relationships in tolKG, extracted data from literature is cross-referenced with structured databases like DisGeNET for validation, with unverified relationships flagged for manual review. Statistical measures such as pointwise mutual information (PMI) help distinguish meaningful associations from coincidental co-occurrences. Additionally, generic or redundant relationships are filtered out, ensuring only well-supported and non-redundant data is integrated, maintaining the reliability and clarity of tolKG. custom dictionaries for tolDC-specific terms, such as tolerising agents and maturation stimuli, were manually curated and integrated into TERMite. Additionally, biomedical ontologies like MeSH and UMLS were utilised to capture broader synonyms and term variations, ensuring comprehensive coverage and improving the accuracy of entity extraction.

5.8.4 Challenges and Limitations of using TERMite

While open-source research practices promote transparency, reproducibility, and broader accessibility, collaboration with commercial partners introduces certain constraints. In this project, we integrated SciBite's TERMite tool, a commercially licensed text-mining solution, to extract biomedical entities from the tolDC literature. Although this collaboration provided access to a robust tool with curated biomedical dictionaries and domain-specific functionality, it also imposed restrictions on what aspects of the project could be shared openly.

One key limitation is that the internal mechanisms of TERMite, such as its proprietary algorithms for named entity recognition (NER) and its curated biomedical vocabularies, cannot be shared openly due to licensing agreements. This means that while the output of TERMite (i.e., the extracted entities) can be included in the tolKG and shared, details of the underlying extraction processes, term-matching logic, and curation methods used by TERMite are not publicly accessible. Additionally, any custom dictionaries provided to SciBite for tolDC-specific terms are considered proprietary extensions within the licensed software and cannot be distributed as part of the open-source outputs.

The restriction on sharing the full extraction pipeline also affects the reproducibility of the work. While open-access users can replicate parts of the process using publicly available NER tools like BioBERT, they cannot fully reproduce TERMite's outputs without a commercial license. Furthermore, code used to integrate TERMite into the text-mining workflow is limited to high-level implementation descriptions, as sharing detailed scripts that interface directly with the proprietary software would violate the licensing terms.

In contrast, structured data sources (e.g., DisGeNET, IntAct, Reactome) used for relationship extraction are freely accessible, and the constructed knowledge graph (tolKG) is openly shared as an RDF dataset. This ensures that the majority of the knowledge graph, particularly its structure and biological insights, remains available to the research community.

5.9 Summary

This is the first study to explore the potential of NER and a graph database for tolDC field. As well as the traditional approach of integrative bioinformatics, using structured data

sources, we have introduced ML-based entity recognition to mine PubMed for additional data. We are now exploring the tolDC field and are able to answer systematic questions that would be hard to answer otherwise. The tolDC field presents significant challenges; there are many terms that are used ambiguously and many entities are referred to with multiple terms. To overcome this, all the entities are standardised in this knowledge graph which makes the comparison possible. The tolDC therapies field has some more specific terms such as cell markers and culture mediums. They can be identified in the NER but not as cell markers or culture medium. For example, the culture medium DMEM was identified as chemical. To solve this, a complete list of cell markers was obtained from cellmarkerdb. All these manual steps increased the efficacy of a tolDC-specific knowledge graph. Currently all the knowledge in the graph, both structured and unstructured is from freely available sources; the use of the NoSQL platform Neo4j should make further expansion simple, including the addition of private data sources if that is needed.

Previously, Literature Based Knowledge Discovery (LBD) has been used extensively to derive molecular interaction in the biomedical field. Some conventional relationship extractions include gene-disease and gene-protein pathways interaction. These relationships are extracted from the entities of published literature, such as Medline abstracts used to identify disease-gene relationships [Kim et al., 2017]. They used a machine learning-based method and concluded that relationship extraction can be improved by improvising the entity extraction. Similarly, the STRING database uses text mining for collecting and scoring the protein-protein interactions [Szklarczyk et al., 2021]. LBD provides benefits over experiment-based knowledge discovery which is mostly expensive and time-consuming. This approach has been applied in many fields, such as Parkinson's, metabolites and is widely studied in new drug discovery field [Kostoff, 2014, Song et al., 2015, Sang et al., 2018]. Mostly, only the abstracts are used for extracting entities from the publications, but the abstracts only provide a summary of the paper and critical information is mentioned in detail in the full text. Moreover, most of the LBD approaches are co-occurrence-based. For example, relationships are built between entities if they frequently appear together in the text. However, some entities can be totally unrelated and can still appear together. A partial solution to this problem has been the use of semantic approaches such as knowledge graphs. After constructing a semantic network, deep learning-based approaches are utilised to build the entities interactions. Such kind of

deep learning-based approaches are also used with knowledge graphs built from biomedical databases. Another example is the data integration from multiple resources to build an ontology for Mitochondrial Disease [Warrender and Lord, 2015]. Traditionally, knowledge discovery is either based solely on literature or the available biomedical databases. In contrast, this study merges data from literature with online data resources.

Despite the successful results, some challenges were faced during the NER and knowledge graph construction. The integration of structured data from web resources enables relationship embedding. The relationship among the entities opens the door to answering and extracting hidden patterns in the data. More relationships can be extracted directly from the text; however, the sentence structure is more complex than in other fields, so we need more advanced techniques to achieve good accuracy. tolKG can also be connected with data analytic tools such as Neo4j Bi Connector to analyse and visualise the data seamlessly. Moreover, we can perform graph mining using ML tools.

Semantic data integration and visualisation are beneficial in the tolDC field for comparing studies, to generating new hypotheses and findings by using already available knowledge. However, there are many issues involved in the tolDC data integration such as weak experimental data reporting, data standardisation and data heterogeneity. While we have managed to circumvent many of these issues during the construction of tolKG, we hope that the existence of this data resource should help to highlight the value of data integration and increase the use of better data practices within the field.

6

Applications of tolKG to enhance the understanding of tolDC therapies

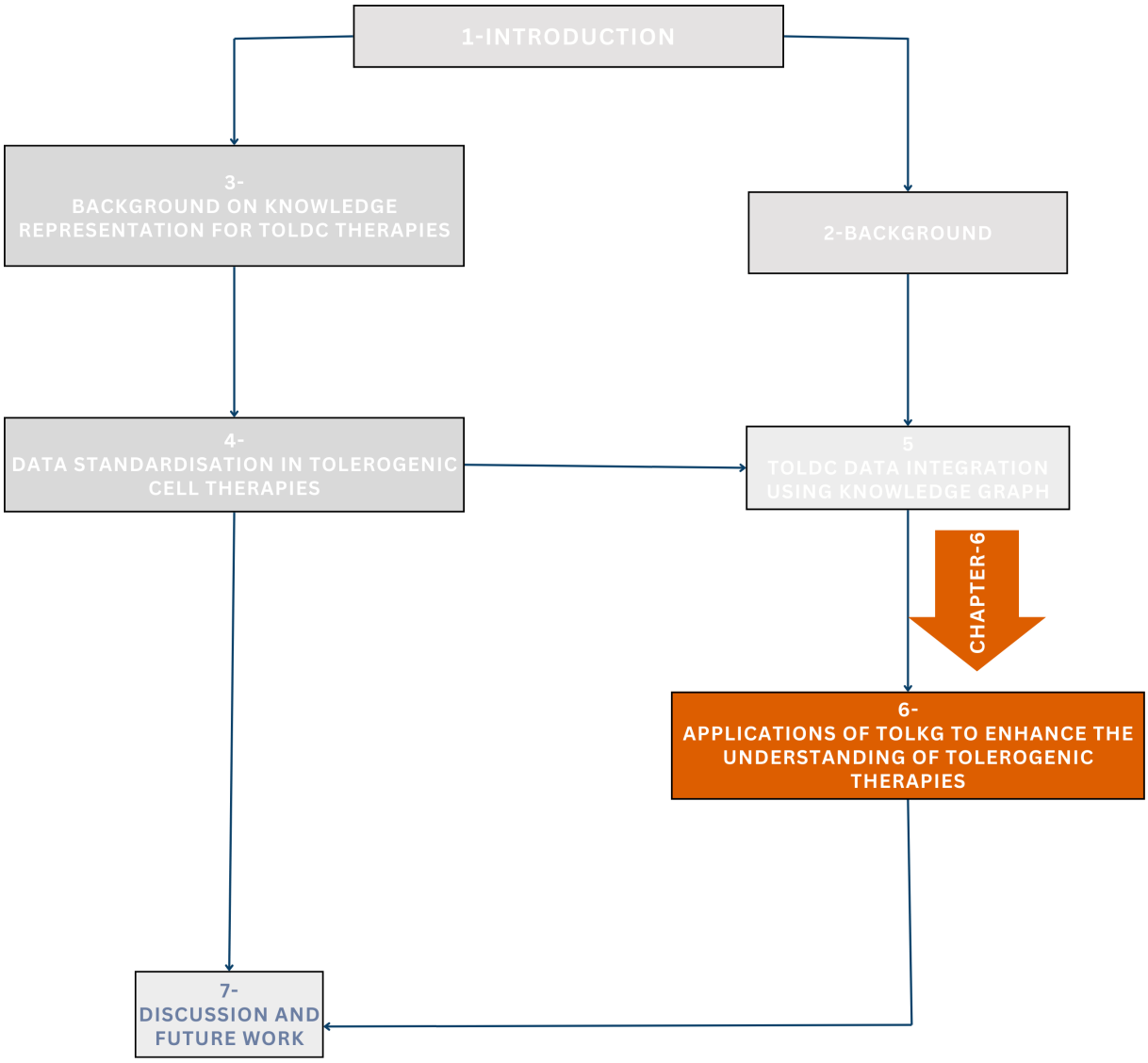


FIGURE 6.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

6.1 Abstract

After constructing tolKG as described in Chapter 5, here we demonstrate the utility of tolKG to understand the tolDC field. The queries are examples of real-world questions that can be answered with tolKG. These queries vary in complexity and also serve as a means to test the overall effectiveness of tolKG.

6.2 Introduction

In Chapter 5, we described a method that enabled the comprehensive and targeted integration of data in the tolDC field. In Chapter 2 and 4, we have highlighted the challenges associated with experimental protocol heterogeneity and lack of understanding about the combined protocols due to standardisation issues. One of the main purposes of building tolKG is to enable us to answer questions related to a set of studies in the tolDC field. In this chapter, we provide examples of the queries that can be performed using tolKG. The queries are designed to test the versatility of the tolKG across various dimensions, from its effectiveness in filtering research studies to its ability to examine heterogeneity in the field.

In Chapter 4, we performed the analysis of a subset of research papers on the tolDC field manually. This analytical endeavour required the efforts of four dedicated researchers who collectively scrutinized more than 20 research papers. The primary objective was to discern and extract the fundamental insights and essential information that the authors of those research papers had provided about the generation of tolDCs. This experiment presented us with a twofold challenge. Firstly, we needed a strategy to identify the pertinent experimental papers within the realm of tolDCs. To address this, we strategically harnessed key terms and specific criteria in our search, aimed at filtering the papers which are focused on the experiments to generate the tolDCs. Secondly, delving into the content of these chosen papers to extract meaningful information required a thorough understanding of the subject matter. Recognizing the intricate nature of the tolDC field, we sought the guidance and expertise of individuals who have degree-level expertise in the tolDC domain. These experts, possessing a profound understanding of tolDC research, were helpful in assisting us with the manual extraction of vital data points and significant findings.

However, amidst these efforts, it became increasingly apparent that the manual nature of this experiment not only demanded substantial time and dedication but also introduced the possibility of oversight or subjective interpretation. In this context, the availability of comprehensive and tailored tools like tolKG can be of significant help. The role of tolKG is to act as a catalyst in simplifying and streamlining these intricate comparative tasks. Through the systematic organisation of diverse knowledge, tolKG has the potential to offer a centralised repository of curated data and insights related to the tolDC field.

The objective of this chapter is to explore the diverse applications of tolKG and improve the understanding of the tolDC field from experimental and social perspectives. To achieve these ends, the chapter is structured to first introduce some basic queries to check if the tolKG is working as it should. Then, we focus on some more in-depth analysis of the field to answer questions related to the heterogeneity of the experiments which is not possible without having a tool like tolKG. In the end, we explore the field from a societal aspect, including examining gender bias within the field.

6.3 Basic Queries for Direct Information: "Does it work?"

In this section, we aim to test the effectiveness and accuracy of tolKG. To do this, we will use simple yet important queries. These queries are essential for checking how well tolKG works in terms of finding and organising information. We have chosen these specific queries because each one will test different parts of tolKG ability to find and process data. Our main goal here is to see if tolKG can really provide reliable and detailed information that is useful for research in the tolDC field.

6.3.1 What are the publications by a specific author?

The first query simply identifies all publications by a specific author in the tolKG system. It serves not only as a direct test of the data extraction capabilities but also as a measure of its precision and reliability in identifying scholarly works. By cross-referencing the results with PubMed, this approach provides a clear benchmark for evaluating the effectiveness of tolKG. This method ensures an objective comparison, highlighting the system's

accuracy in aligning with established repositories and its potential use in streamlining tolDC research.

We have chosen to search for all works by "Professor Catharien Hilkens"; we have chosen this name as it is relatively unusual and is likely to produce only papers by a single author. This is appropriate because the query is simply testing the workings of the knowledge base; clearly, for a more general solution identifying an individual scientist, we would need a more complex, heuristic approach, as, despite the development of Orcid IDs, these are often not given in PubMed.

We executed a simple query in tolKG using Cypher to gather all publications authored by "Professor Catharien Hilkens" and compared the total with those listed on PubMed. When searching for publications by a particular author, it is important to note that the name of the author might be listed differently across various works. For instance, "Professor Catharien Hilkens" has publications under "Catharien M U Hilkens","Catharien Mu Hilkens" or "Catharien M.U Hilkens". To address these variations, we employ a strategy in our Cypher query that includes using the *CONTAINS* keyword. This allows us to refine our search to encompass records where the first name contains "Catharien" and the last name contains "Hilkens". This approach helps in capturing the broad spectrum of name variations under which their work may be published.

Our tolKG search successfully identified 34 publications authored by the specified individual (March 2024). To verify the comprehensiveness of this retrieval, we conducted a parallel search on PubMed, specifically filtering for open-access articles, which yielded 41 publications by the same author. Upon comparing the PMC IDs from the tolKG results against those found on PubMed, we noted a discrepancy of 6 papers. These missing documents were published after the creation of tolKG. This comparison confirms the accurate integration of data within tolKG and validates its effectiveness in fetching relevant papers.

6.3.2 What are all the method papers of the tolDC field?

The strategic categorisation of research papers plays a pivotal role in various research tasks, particularly in tasks like building comprehensive corpora for meta-analysis and conducting exhaustive literature reviews. A practical application of this approach was encountered in Chapter 4, where our objective was to identify all methodology papers in

the tolDC field to analyse the prevalence of the MITAP application. This task, however, was not straightforward. As a result, a series of complex search queries in PubMed were employed to gather the necessary data. It is possible to achieve this query with tolKG because it contains a categorisation of papers based on data from PubMed as described in Section 5.5.3.

This preliminary query, "List all the method papers of the tolDC field", serves a dual purpose in evaluating the tolKG system. Firstly, it aims to test the fundamental data retrieval capabilities of tolKG, focusing on its ability to accurately extract and list specific categories of scientific literature. Secondly, this query assesses the comprehensiveness of literature integration in tolKG, as a well-functioning knowledge graph should encompass a wide array of methodological papers relevant to the tolDC field.

Evaluation

Unlike our previous query, looking at authors, we lacked a gold standard result to test this query. Therefore, we used our prior work on MITAP, which resulted in a manually curated set of these papers as described in Chapter 4. To allow us to compare directly to this dataset, we used Cypher command in tolKG to extract protocol papers published post-August 2016, which is the publication date of MITAP. Given that the MITAP analysis was conducted in March 2021, our tolKG query spanned from August 2016 to March 2021. This search yielded a total of 152 research papers, a number surpassing our manual search results on PubMed, as referenced in Section 4.5.2.

We wished to assess the precision and recall of these 152 papers. The recall can be assessed by considering how many of the 72 manually sourced papers found in our original MITAP analysis were also present in the tolKG query; we found that 67 of the 72 papers were present in the tolKG dataset. Of the 5 that are missing, four of these were not open access which explains their absence from the tolKG results. The 72 papers that we are looking for in the tolKG retrieved results are also retrieved by a PubMed query which has a precision and recall of 75 and 83 percent respectively, which means it is not 100 percent accurate. However, the approach we used in Section 5.5.3 showed a higher recall rate than the PubMed strategy discussed in Section 4.5.2, indicating it not only covered all the papers identified by the PubMed query but also identified additional relevant papers.

To assess the relevance of these additional papers, we examined the precision by reviewing a random selection of 10 papers from the tolKG-sourced papers, explicitly excluding the initially identified 72 papers. This review revealed that 8 out of the 10 were relevant

MITAP papers, resulting in an estimated precision of 80%, surpassing the PubMed query utilised in Section 4.5.2. Although this is not 100% precise, it is considered more effective than the customized manual query.

This demonstrates the capability of tolKG to produce a detailed and more accurate compilation of methodological papers, highlighting its value as a research tool. This efficiency in gathering literature not only enhances the review process but also aids in integrating knowledge within the tolDC field.

Ultimately, these findings suggest that tolKG is operating effectively, enabling comprehensive searches that were previously unattainable, thus affirming its utility in advancing research in the tolDC domain.

6.4 Re-examining the adoption rate and impact of MITAP

Why this query is important?

In the analysis of MITAP, detailed in Chapter 4, a pivotal step involves determining the number of papers that might have employed MITAP. This assessment was conducted using a refined keyword search strategy on PubMed. After evaluating both precision and recall, the final query yielded an approximate count of 72 papers that could have incorporated MITAP. Notably, during this analysis, we lacked a tool like tolKG to assist in this inquiry. Now, we revisit the question: post its publication, how many papers could have used MITAP?

MITAP's impact based on tolKG During the MITAP analysis, it was found that only 14% of researchers are using MITAP. This was based on the manual query on Pubmed. The analysis was performed in March 2021 and MITAP had 40 citations at that time, out of which 10 were actually using MITAP. Using tolKG to retrieve the method papers post-August 2017 to March 2021 retrieves 122 papers as shown in Figure 6.2. It reduces the impact OF MITAP further to 8%. Moreover, we rechecked the citations of MITAP and there are 2 more papers now that are using MITAP, which makes the total count to 12. So the latest MITAP impact, as per September 2023, is 9%.

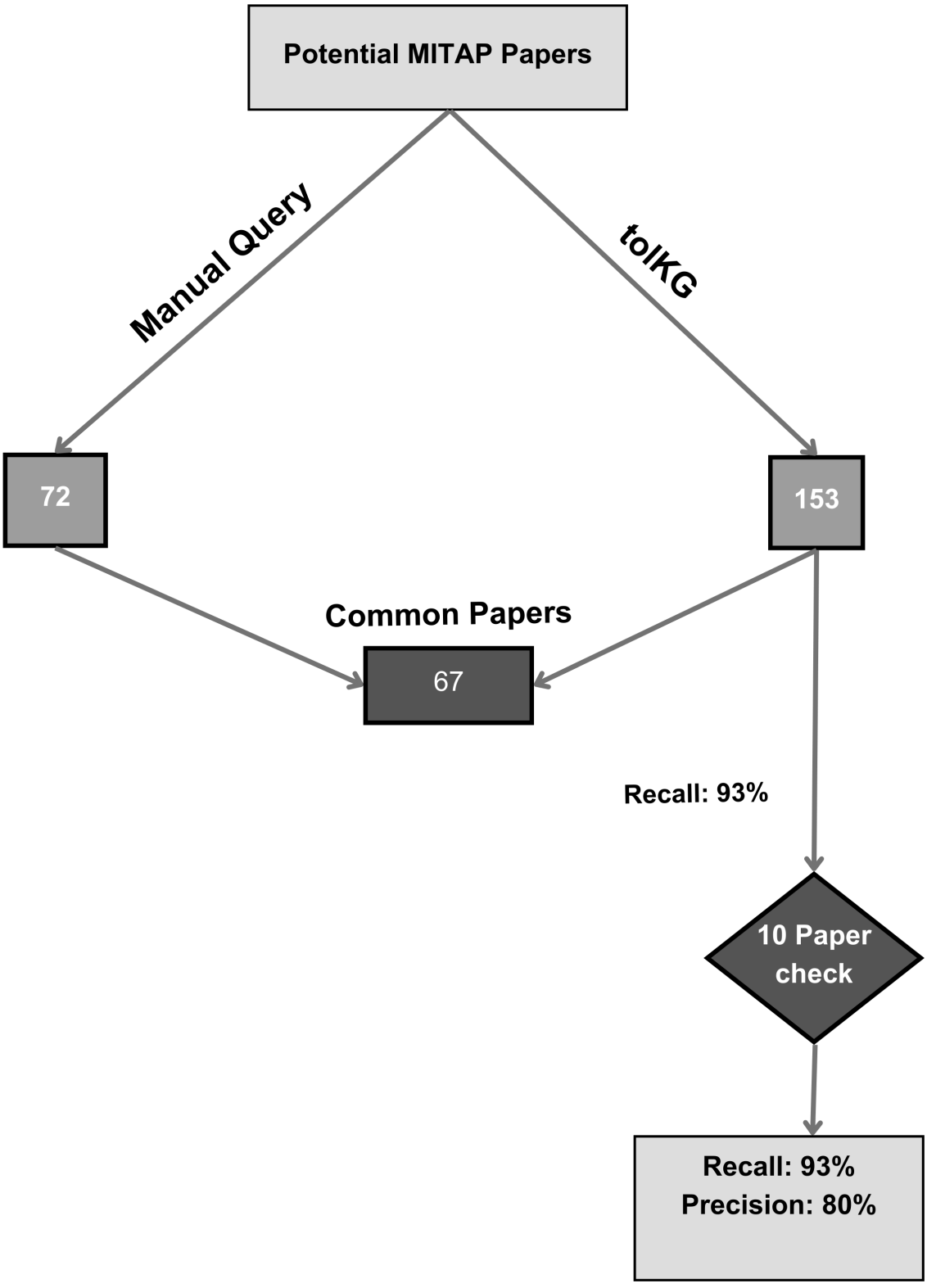


Figure 6.2: Potential MITAP papers with tolKG
Analysis of recall and precision of potential MITAP papers by using tolKG

How using tolKG could have made our life easier?

Using tolKG can save time, enabling researchers to make progress more swiftly. Instead of analysing a vast array of papers, the narrowed focus allows for a more efficient allocation of efforts, ensuring that the analysis is conducted on papers directly relevant to the specific category of interest.

This query also shows tolKG's potential to significantly enhance the efficiency of literature review processes and knowledge integration in the tolDC domain. Firstly, it aims to test the fundamental data retrieval capabilities of tolKG, focusing on its ability to accurately extract and list specific categories of scientific literature. Secondly, this query assesses the comprehensiveness of literature integration in tolKG, as a well-functioning knowledge graph should encompass a wide array of methodological papers relevant to the tolDC field.

6.5 Understanding heterogeneity in the tolDC field

The tolDC field experiences data variation due to many different experimental protocols, standardisation issues and analytical methods (see Section 3.2.4). In this section, we attempt to quantify the extent of this heterogeneity: are protocols widely varied in terms of reagents and procedural steps; or is there a commonality across most research publications? We aim to explore the range of experimental designs and cell preparation techniques within tolDC therapy research using tolKG to understand the variability and uniformity in approaches. Here, our goal is to understand the diversity in tolDC field by exploring common practices and preferences within the field. This not only underpins experimental design but also contributes to the broader goal of developing immunotherapeutic interventions. For example, knowledge about commonly used methodologies helps in standardising experimental protocols across the field which is vital for comparing results and advancing collective understanding. It also facilitates replication of studies and validation of findings which is an essential step for scientific progress. Using well-established and effective methodologies can enhance the safety and efficacy profiles of tolDC therapies for translational and clinical applications. For example, by identifying the most commonly utilised methodologies, we can know about which tolDC therapies may soon transition into clinical practice. Furthermore, the collective preference for certain methodologies can guide future research, spotlighting areas for innovation or underscoring the

need for alternative strategies to address the limitations of current practices. Ultimately, cataloguing the preferred methodologies in tolDC research charts the progress within this specialised area and also propels the entire field towards more effective, standardised and translatable therapies.

MITAP implicitly suggested which sources are important to be described in tolDC research. These are the sources that have a significant effect on the experimental outcomes and are crucial to be reported for reproducing the results. We followed MITAP to identify which aspects of heterogeneity in tolDC research are examinable through literature. The general steps to generate tolDCs are described in Section 3.1. Here, we will evaluate the heterogeneity in the process of producing tolDCs step by step.

6.5.1 What are the agents to induce the tolDCs?

Agents used to induce tolDCs include a variety of pharmacological and biological substances that modulate dendritic cell function towards a more regulatory or tolerogenic state. We care about this because different agents have different modes of practice and could produce functionally different cells. In addition, different agents normally have distinct protocols.

The protocols for generating tolDCs differ based on the agents or methods used. Some of the most common agents include:

- Vitamin D3: 1,25-dihydroxyvitamin D3 modulates the immune system by binding to the Vitamin D receptor (VDR) on dendritic cells. It reduces the expression of co-stimulatory molecules and increases the production of anti-inflammatory cytokines like IL-10.
- Dexamethasone: Dexamethasone acts through the glucocorticoid receptor to suppress inflammation and immune responses.
- Rapamycin (RAPA): A mammalian target of rapamycin (mTOR) inhibitor can decrease the production of pro-inflammatory cytokines and increase the production of IL-10.

Although these agents vary in their mechanisms, their primary goal is to enhance tolerogenic characteristics while minimising inflammatory responses. We know from hearsay

that there is significant variation in choice of agent, but here we aim to quantify this variation; this would enable us to understand how significant this variation is as a barrier to comparability between different researchers.

How we did this?

1. In tolKG, original methodology papers that use protocols to generate tolDCs are categorized, totalling 5,963 papers.

2. From these, the method sections of 3,509 papers were extracted as outlined in the methodology discussion.

3. To examine the agents used in the protocols, we counted the mentions of the agent names. In the full original papers, the count limit was set to 3 and in the method sections, the count limit was set to at least 1. This limit was chosen based on careful observation after reviewing a subset of papers rather than being entirely arbitrary.

	All Protocol papers (6k)	**Also including methods (3.5k)**
VIT-D3	630 (10.5%)	180 (5.1%)
DEXA	1070 (17.8%)	470 (13.4%)
RAPA	1140 (19.0%)	550 (15.7%)

Table 6.1: Counts and percentages of paper types for VIT-D3, RAPA, and DEXA

RAPA appears to be the most frequently discussed agent in both protocol and methods-only papers, indicating a possibly higher level of research activity or interest related to RAPA compared to VIT-D3 and DEXA. However, RAPA is far from universal as DEXA and VIT-D3 are both common. It indicates a high level of interest or research activity around Rapamycin, which could be due to various factors including its effectiveness, availability or how well its mechanisms are understood. The effectiveness of an inducing agent

should be evaluated based on specific research outcomes, safety profiles and the context of its application.

6.5.2 How many different kinds of growth media are there and what are the most frequently used?

Once inducing agents have been chosen, next we consider the culture media which is a crucial component for effectively inducing and maintaining the immunosuppressive and tolerogenic qualities in tolDCs. The culture media provides a vital environment, offering the necessary nutrients, growth factors and physiological conditions that support cell survival and functionality. To examine the heterogeneity in the tolDC field, knowing the different types and popular growth mediums for growing tolDCs is important.

With tolKG, we wanted to see what are some most frequently used culture media. It is important to note that the composition of culture media can vary depending on the specific protocols, research goals and cell types being cultured. Researchers may modify or supplement these culture media with additional factors or components to enhance the generation, maintenance and functionality of tolDCs.

In tolKG, we have categorised the method papers, yet determining the actual use of culture media from mere mentions within these papers is challenging. That is why, we filtered out the methods sections to identify specific culture media references. We assume their usage in the experiments based on these mentions.

Although this approach helps in understanding the diversity in cultural media selection, the accuracy of these findings may not be absolute, serving primarily as an indication of the range of preferences and practices.

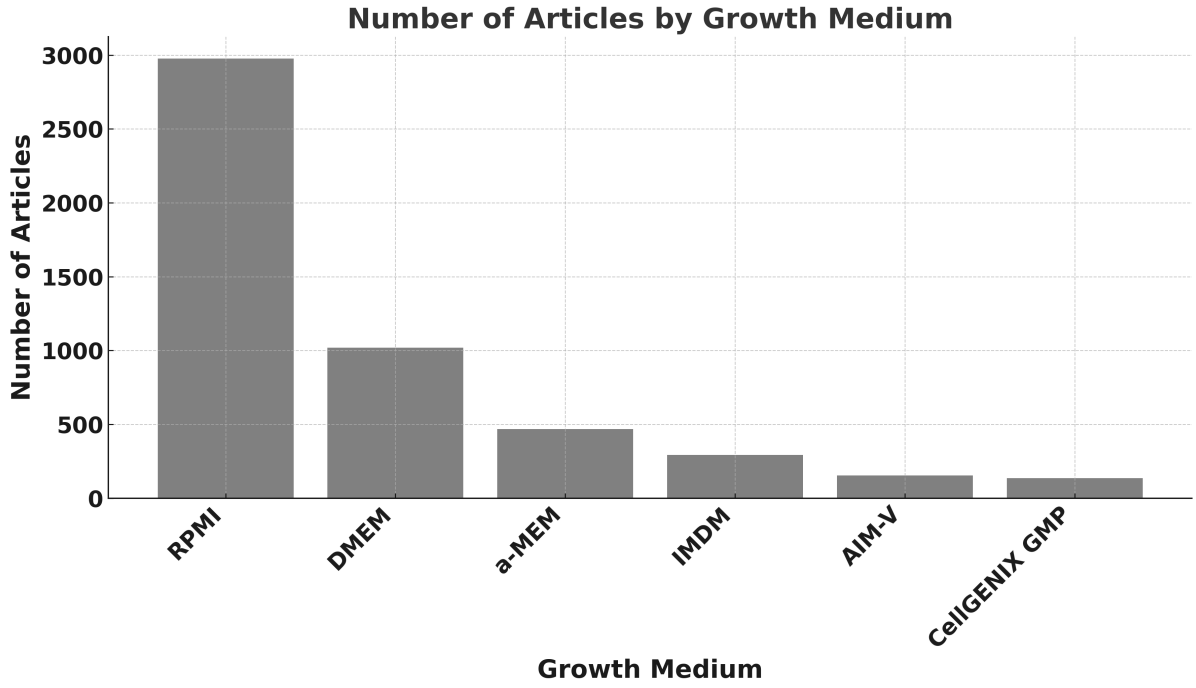


Figure 6.3: The graph shows the top most used growth mediums in the tolDC field

RPMI, alpha MEM, DMEM and IMDM are common basal media that, depending on the specific requirements, might be supplemented with factors such as serum, cytokines or antibiotics to support tolDC development. AIM-V and CellGenix GMP DC media are specialised, serum-free options designed to meet the stringent requirements for clinical applications, providing a controlled environment for tolDC cultivation with a focus on safety and efficacy in therapeutic settings. GMP is a standard term that stands for good manufacturing practice and describes the minimum standard that a medicine manufacturer must meet in their production processes.

The preference for RPMI over CellGenix GMP in the tolDC field likely reflects the versatility and established efficacy of RPMI in dendritic cells. It is a well-characterised medium that offers researchers flexibility in customisation for specific experimental needs by adding supplements. CellGenix GMP, while specifically designed for clinical-grade cell production, including tolDCs, might be less utilised due to its higher cost and specificity, making RPMI a more accessible choice for routine research applications.

6.5.3 How are tolDCs matured?

The maturation process in dendritic cells refers to their transition from an **immature state**, where they efficiently capture antigens, to a **mature state**, where they specialize in presenting these antigens to T cells and triggering an immune response. This process is crucial for the activation of the immune system. A central question in the study of tolDCs concerns whether these cells undergo maturation and, if so, the specific factors employed to achieve this state. The maturation of tolDCs can be influenced by various stimuli, with lipopolysaccharide (LPS), monophosphoryl lipid A (MPLA) and a cocktail of cytokines being among the primary agents investigated. LPS, a component derived from the outer membrane of Gram-negative bacteria, is known for its potent ability to activate dendritic cells. MPLA, a detoxified derivative of LPS, serves as a safer alternative, still capable of inducing maturation but with reduced toxicity. Alternatively, a combination of cytokines can be used to tailor the maturation process, potentially offering a more controlled and nuanced approach to inducing the desired tolerogenic state. Understanding which of these factors is used and under what circumstances is crucial for refining tolDC-based therapies, aiming to maximize their therapeutic potential while minimizing adverse effects.

The question "Were tolDC matured and if so, which maturation factors were used?" seeks to determine whether tolDC underwent a maturation process in a given study or experimental setup and if so, it asks for specifics on the maturation factors or agents employed to induce this process. This inquiry is relevant for understanding the methods used to prepare tolDC for therapeutic applications, as maturation status can significantly influence their immunological properties and effectiveness.

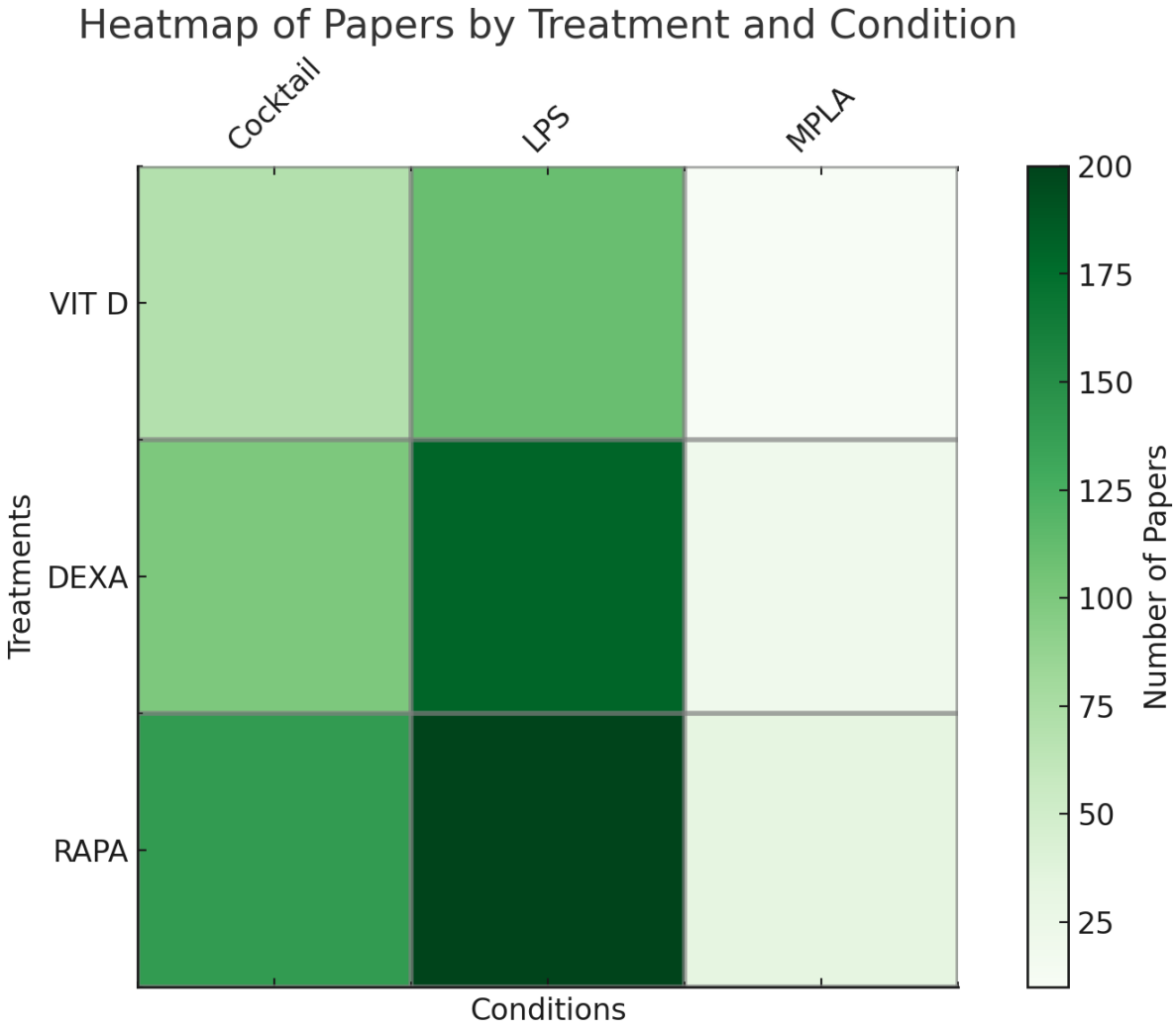


Figure 6.4: The graph shows usage of tolerogenic (VitD3, Dexa and Rapamycin) and maturation agents in the tolDC field. Here 'cocktail' entails (cytokine cocktails)

The variation in tolDC for tolerogenic agents—Vitamin D3, Dexamethasone and Rapamycin, reveals distinct experimental strategies. Vitamin D3 is frequently combined with cytokine cocktails or LPS, indicating its role in modulating inflammatory responses. Dexamethasone, known for its anti-inflammatory effects, is predominantly used with LPS, suggesting its effectiveness in reducing inflammation in tolDC therapies. Rapamycin shows diverse application patterns, particularly with cytokine cocktails and LPS, highlighting its significance in promoting a tolerogenic phenotype. The absence of combinations with MPLA for all three agents points to specific research preferences in the field.

The absence of MPLA use in certain studies could be due to various factors, including the preference for other maturation agents with more established or desired effects, potential regulatory or availability issues or specific research focus areas that align better with the mechanisms of action of Vitamin D3, Dexamethasone or Rapamycin. Each agent has unique properties and impacts on tolDC maturation and researchers might choose based on the specific immunological outcomes they aim to investigate.

6.6 Do all tolDC protocols relate to a similar category of diseases?

One of the intriguing questions in the field of tolDC therapies is the existence and roles of different types of tolDC. Our exploration focuses on understanding the diversity among tolDC types and examining whether their functions are uniform or distinct.

6.6.1 What are the disease networks of deferentially expressed genes of tolDCs generated using different protocols?

As discussed before, tolDC therapies are generated using a variety of protocols. An important question to consider is whether different protocols for generating tolDCs have a different impact on the expression of genes associated with specific diseases. It is possible that some protocols may work better or have negative side effects for specific diseases. This inquiry builds upon a previous study that examined differential gene expression of tolDC induced with vitamin D3 (vitD3-tolDC), dexamethasone (dexa-tolDC), or rapamycin (rapa-tolDC) using microarray analysis in 5 healthy donors [Navarro-Barriuso et al.,

2018]. The results revealed that no common differentially expressed genes (DEGs) were identified across all three tolDC protocols. Nevertheless, the study identified individual genes that may serve as biomarkers for each protocol. Moreover, a gene set enrichment analysis indicated that dexa-tolDC and vitD3-tolDC share several immune regulatory and anti-inflammatory pathways, whereas rapa-tolDC seems to induce tolerance through a robust suppression of cellular processes.

Currently, tolDCs are being investigated as a potential therapeutic strategy for various immune-related diseases, such as diabetes, rheumatoid arthritis, multiple sclerosis and inflammatory bowel disease. In this study, we sought to investigate the gene-disease networks of DEGs in the three types of tolDCs to determine whether the signatures associated with each of the three tolDC protocols are related to similar or distinct diseases. Our objective was to identify the genes within each DEG set that are related to one another through at least one disease, thus providing a defined set of diseases that each of these DEG sets represents. To achieve this, we employed the Entrez IDs of the DEGs as input for tolKG and formulated a cypher query to retrieve the relevant disease associations.

The present analysis revealed that the disease-associated networks of DEGs of dexa-tolDCs and vitD3-tolDCs are related to similar disease categories. However, the disease-associated network of rapa-tolDCs DEGs appears to be distinct, particularly with the inclusion of depression, thalassemia and insomnia which are not represented by DEGs of the other protocols as ca be seen in Figure 6.5. This finding further supports the hypothesis outlined in the original study that rapa-tolDCs may use a distinct mechanism of tolerance induction.

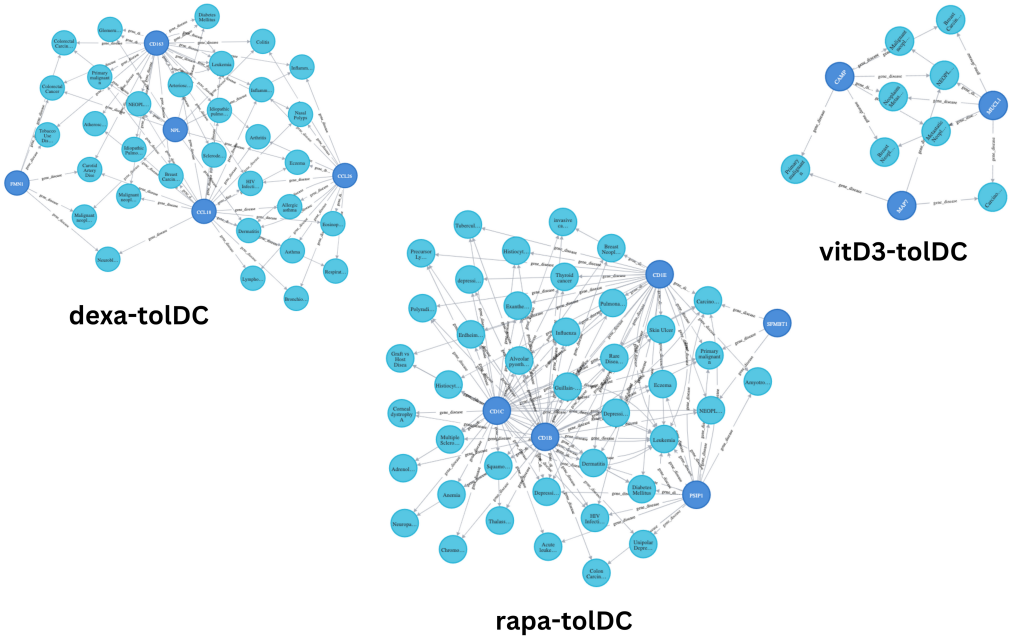


Figure 6.5: The overall result of NER and structured data integration into the Neo4j graph database.

While vitamin D3 and dexamethasone-induced tolDCs share some immune regulatory and anti-inflammatory pathways, rapamycin-induced tolDCs operate through a different mechanism, mainly involving strong immunosuppression. This indicates that although these various types of tolDCs are all aimed at inducing immune tolerance, they do so through different molecular pathways and mechanisms.

6.6.2 Do Rapa-tolDCs and other tolDCs target different diseases?

Section 6.6.1 revealed a direct association between depressive disorder-related diseases and tolDC therapies derived from rapamycin, indicating the potential of such therapies to address these conditions. However, these categories of disease are missing from the DEGs of the other two types of tolDCs.

To further explore this association, we extended our analysis to investigate if any of these diseases were indirectly associated with the DEGs of vitD3 or dexa-tolDCs using graph traversal queries in Neo4J. Specifically, we utilised the shortest path search in cypher to identify any immediate, indirect link between *Depressive Disorder* and the DEGs of vitD3-tolDCs and DEXA-tolDCs.

The result demonstrated that the disease *Depressive Disorder* is linked with the disease *Neoplasm* which is a prominent disease present in the DEGs associated with vitD3-tolDCs and dexa-tolDCs. *Neoplasm* refers to an abnormal mass of tissue resulting from excessive cell growth or failure to undergo programmed cell death, which can be either benign or cancerous. *Depressive disorder* is a common psychological reaction to a cancer diagnosis and studies have shown that individuals with cancers are more likely to experience depressive symptoms compared to the general population, with up to 25% of cancer patients experiencing depression [Niedzwiedz et al., 2019]

This example highlights how graph traversal queries can be used to investigate relationships between derived concepts, such as disease-disease associations. Moreover, the analysis reveals that although *Depressive Disorder* is not directly linked with the DEGs of vitD3-tolDCs and dexa-tolDCs, there exists an indirect path between them through other disease categories.

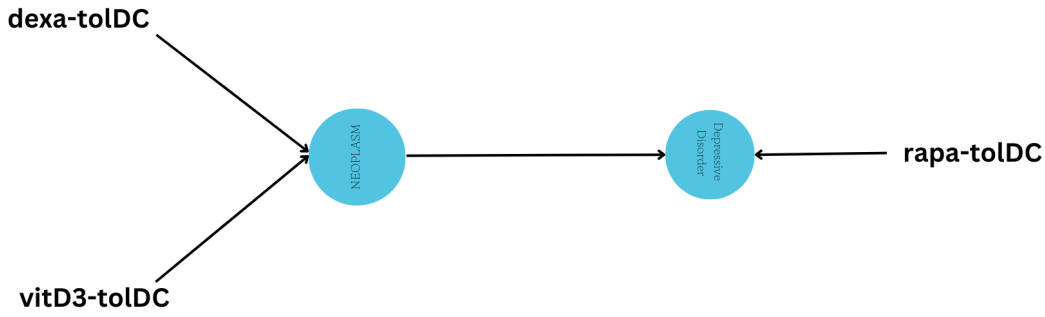


Figure 6.6: Shortest path between specific diseases such as depression which is only related directly with rapa-tolDCs and DEGs of Dexa-tolDCs and vitD3-tolDCs

The indirect associations revealed through graph traversal queries, particularly the link between depressive disorders and diseases like neoplasms, which are present in the DEGs of vitD3-tolDCs and dexa-tolDCs, underscore a shared functional aspect of these different tolDC types in impacting complex disease processes.

6.7 Social Environment Analysis

Here, we perform an analysis of the gender bias in the tolDC therapies field. Our comparative approach seeks to analyse whether there are any notable differences in authorship representation based on gender and nationality across various publications related to tolerogenic therapy. Such an investigation is timely, as there has been a growing emphasis on fostering inclusivity and diversity in the scientific realm. By examining the gender of authors, the study can unveil potential gender-based imbalances in participation and recognition within the field. Similarly, the assessment of author nationality could provide insights into global collaboration dynamics, potential biases and the distribution of expertise across different regions. This research holds the potential to highlight any disparities that exist, enabling the scientific community to address possible inequalities and work towards fostering a more equitable research environment. Additionally, the findings could prompt discussions on the factors that contribute to these patterns and inspire initiatives aimed at promoting diversity, inclusivity and international collaboration in the field of tolerogenic therapy and beyond. Ultimately, this comparative study contributes to the broader conversation on the importance of representation and inclusivity in scientific research, advancing the understanding of authorship dynamics and their implications within a specific domain.

In Figure 6.7, we present the analysis of gender bias among first authors. The decision to have the first author as the main analyst was intentional because the first author typically takes the lead in writing the paper, compared to the other authors. As explained in the methods section, we extracted the full names of all authors from XML files. Since the number of authors varies from paper to paper, identifying the first author was the simplest and most consistent approach. It was much easier to access and analyse the first author's details than to track other authors, such as the last author or those in between, whose positions change depending on the paper.

The chart illustrates that in many countries, male first authors dominate. Notably, the

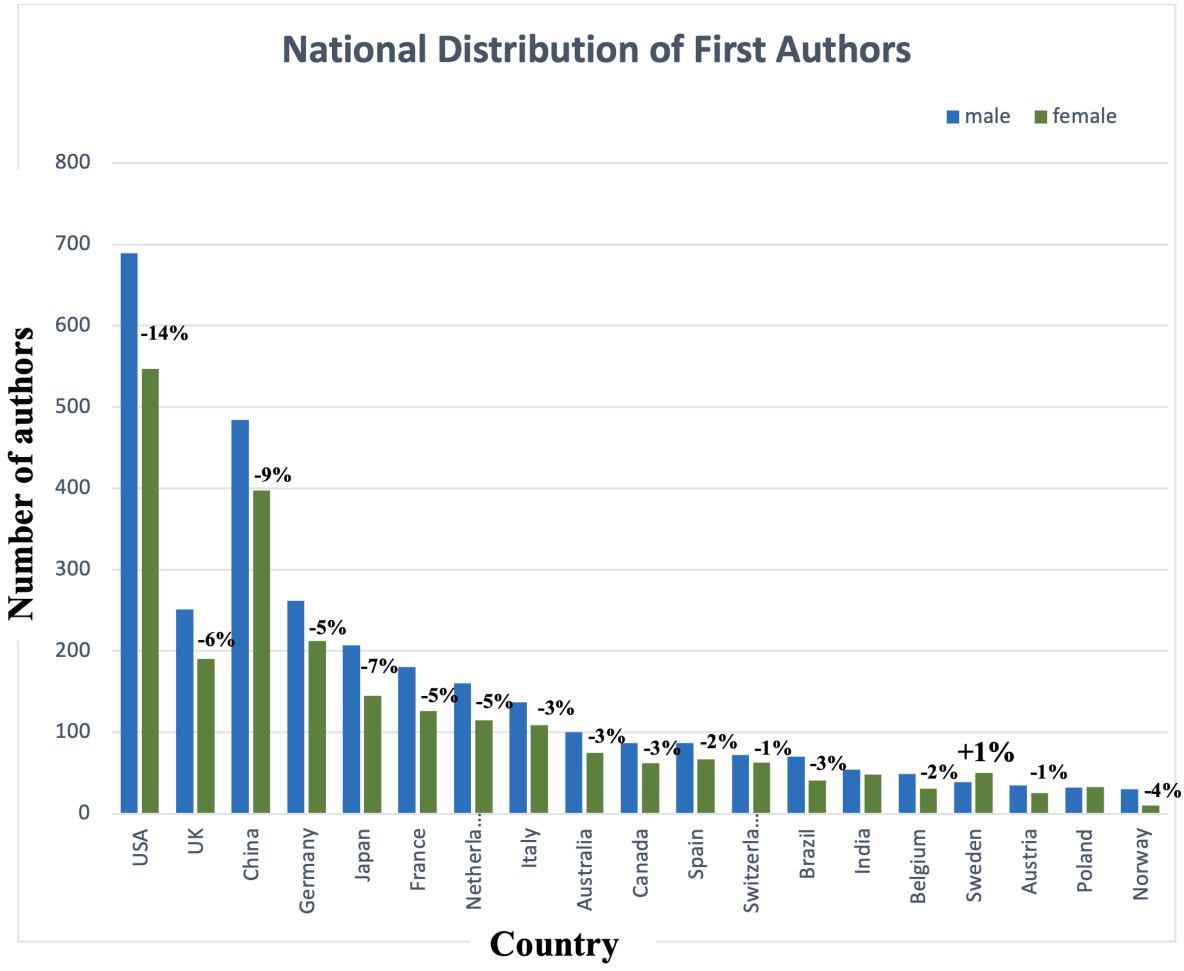


Figure 6.7: Layout of structured data integration into tolKG by using APIs of IntAct, DisGeNET and DGidb.

most pronounced difference of 14% is observed in the USA. This disparity could potentially be attributed to the larger population size of the USA and a greater number of universities compared to other regions. Sweden, however, only exhibits a marginal difference of 1% more female first authors than males, which is not statistically significant. Overall, the distribution of first authors in the field of tolerogenic therapy skews towards males.

Recent studies have highlighted persistent gender disparities in academic authorship, with women remaining underrepresented as first and senior authors and facing limited opportunities in high-impact journals [Bendels et al., 2018]. This study aligns with these findings, confirming ongoing gender imbalances in authorship.

Unlike many studies that rely on automated gender inference, this research employs a more precise method by extracting full author names from XML files, reducing misclassification risks. Despite this difference, the results reinforce existing literature, providing further evidence of gendered authorship patterns and the structural barriers affecting female and gender-diverse researchers.

6.8 Conclusion

The analysis illustrates a clear preference within the research community for combining these immunomodulatory agents with LPS, possibly due to the well-documented role of LPS in dendritic cell maturation and activation. The frequent use of cytokine cocktails across all three agents reflects a strategic approach to emulate the complex cytokine environment encountered by dendritic cells in vivo. The complete absence of studies exploring the combination of these agents with MPLA suggests either a lack of interest or potential challenges in demonstrating efficacy or relevance.

This data highlights areas of concentrated research effort and potential gaps that could guide future investigations. The notable absence of MPLA combinations across all agents points to an opportunity for novel research, exploring whether these agents could synergize with MPLA to enhance the tolerogenic potential of dendritic cells further. Additionally, the varying degrees of focus on different agents with LPS and cytokine cocktails indicate a nuanced understanding of how each agent affects tolDC maturation and function, offering valuable insights for the optimization of tolDC therapies.

This chapter illustrates the profound applications of tolKG in advancing the understanding of tolDC therapies. Through diverse queries, from basic verifications to complex inquiries into heterogeneity and societal impacts, tolKG has demonstrated its utility in understanding the tolDC field. The findings also highlight the importance of standardised methodologies, as evidenced by the MITAP analysis and emphasizes significant variations in tolDC maturation practices. Furthermore, the exploration into the social environment of research unveils the necessity for increased diversity and collaboration. Collectively, these insights not only contribute to the scientific dialogue surrounding tolDC therapies but also paves the way for future research, emphasizing the need for a more inclusive and comprehensive approach in scientific endeavours.

7

Discussion and Future Work

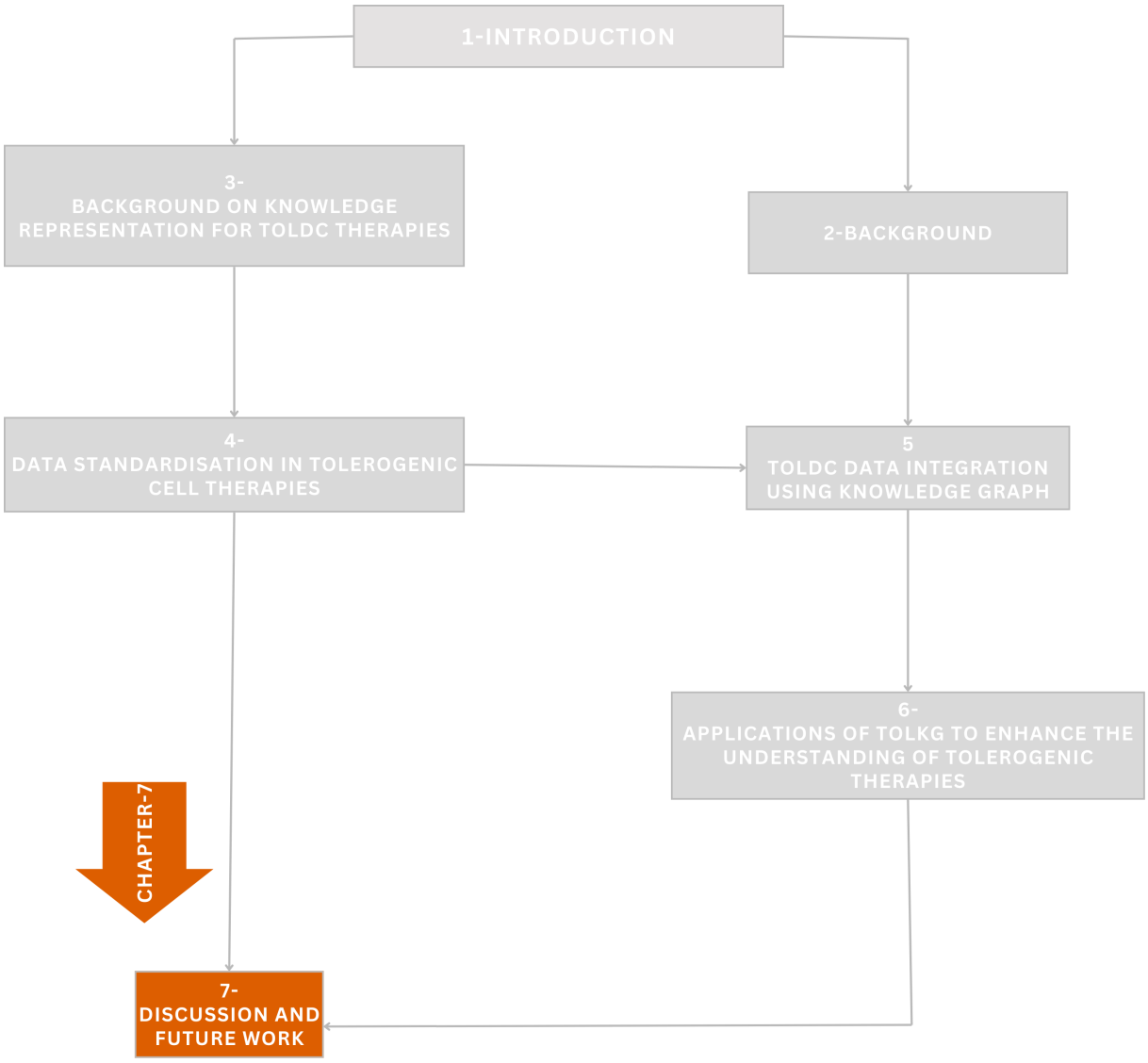


FIGURE 7.1: Layout of the thesis
Overview of the thesis structure, illustrating the logical flow from background research to data integration and evaluation.

7.1 Abstract

In this chapter, we aim to discuss the outcomes and key findings of the project in the context of the initial research objectives. We discuss the key findings of each objective in terms of their contribution to the field. Then, we discuss the limitations of the work. We also identify and introduce future opportunities for this work. In the end, we discuss the broader context which includes some unique insights and reflections on the research conducted.

7.2 Introduction

In this thesis, we describe the methodologies that enable data repurposing and knowledge representation in relatively new fields of biology. We identified that there are two major requirements to practically implement such strategies which are data standardisation and data integration. During this project, an extensive literature review is conducted to find the impact of Minimum Information Models (MIMs) on a subfield called tolDC therapies, see Chapter 4. We also had some interesting findings and ideas to promote the usage of MIMs overall. In addition, we developed a framework to perform a targeted and comprehensive data integration, called *tolKG*. This data integration involves several important steps and data mining from the literature as described in Chapter 5. The purpose of tolKG is to resolve the heterogeneity issues in the tolDC field as discussed in Chapter 2. These include several problems varying in complexity and explain the real-world applicability of the tolKG, Chapter 6. Furthermore, tolKG is shown to be useful in extracting the implicit knowledge from the combined datasets and can help researchers build new hypotheses to accelerate the development of tolDC therapies.

In Section 1.1, we stated that we define the data, information, knowledge and wisdom according to the DIKW pyramid (see Figure 1.2). In Chapter 5 and Chapter 6, we took a series of steps that transformed the data into information, knowledge and wisdom. In Figure 7.2, we provide an illustration of the application of DIKW pyramid to tolKG, which demonstrates how stepwise structuring, contextualisation and integration of graph-oriented data can be used to drive insights. The process begins with raw text data from research publications. In its initial form, this data is difficult to analyse or connect with other relevant work because it is not in a machine-readable format. In the next step, we transformed this data about the tolDCs into a linked knowledge graph where we can easily perform comprehensive analysis with other research work as well as how the entities are connected to each other within a research paper. The figure demonstrates an example of the work that we explained in Section 6.6, where we performed an analysis of an experimental study. The third step where this information is converted into knowledge is when we take the entities from this specific research paper and learn new relationships across domains. Based on this knowledge, we were able to draw a hypothesis that although different types of tolDCs have different pathways, eventually, they are linked to the same category of diseases indirectly. This actionable knowledge is categorised as wisdom.

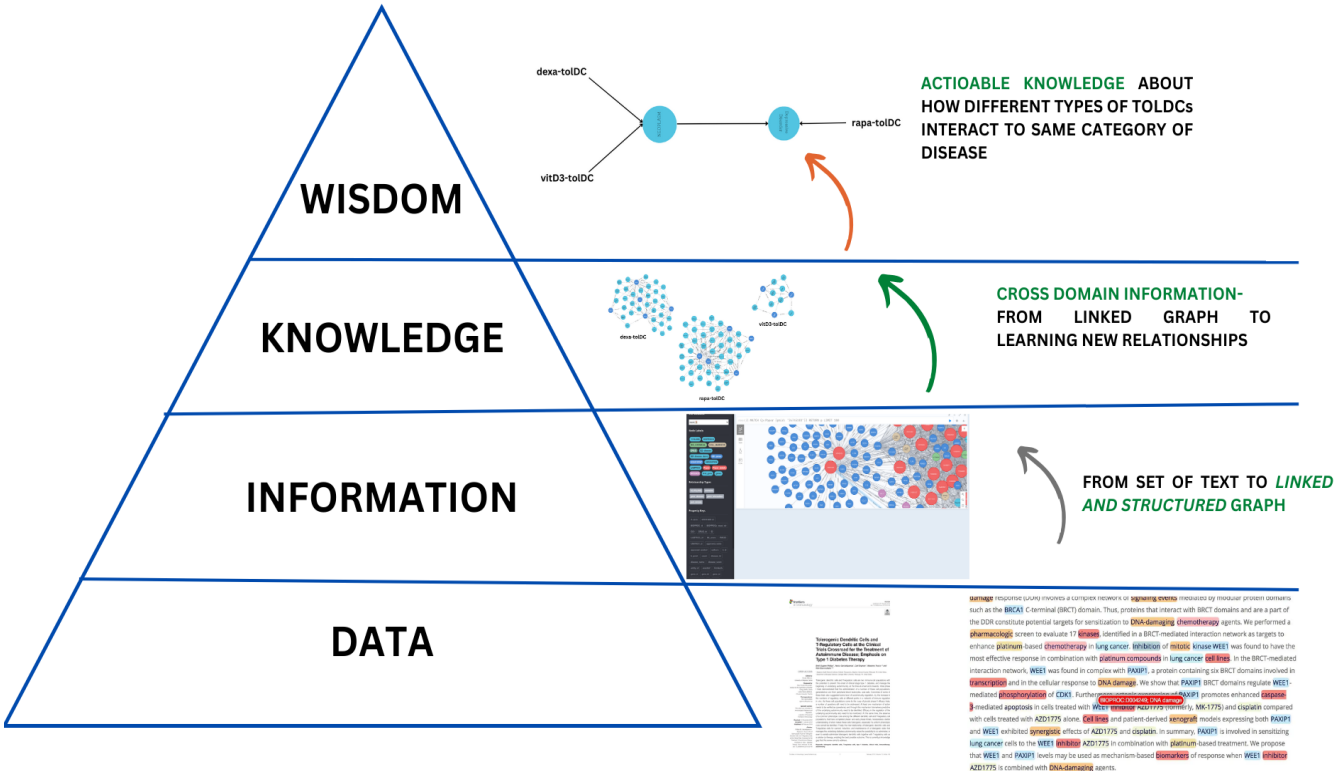


Figure 7.2: tolKG DIKW
An illustration of the application of DIKW pyramid to tolKG.

Furthermore, in this chapter, we describe the outcomes of the thesis in terms of the preliminary objectives set in Chapter 1. For each research question, we assess the thesis outcome in terms of their contribution to the field. Then, we discuss the limitations of the approaches described. We also assess the project outcomes in the broader context and of the recent developments that happened in the field. Finally, we identify and explain the future directions and opportunities.

7.3 Discussion of key findings/ revisiting the main objectives

To fulfil the main aim of this project, three research objectives were set at the beginning of the project (as described in Section 1.3). These are:

1. To improve and promote data standardisation in the tolerogenic dendritic cell field.
2. To integrate the already available data into a database
3. To understand the heterogeneity in the tolDC field

In the next sections, we explain how we have resolved these research questions.

7.3.1 Improving and promoting standardisation in the tolerogenic dendritic cell field

We performed an extensive analysis of MITAP to improve and promote the standardisation in the tolDC field (see Chapter 4). This analysis involved assessing the impact of MITAP on the field after almost five years of its publication, as discussed in 4.5. We found that MITAP is followed by only a small percentage of research, which is 14%. This highlights the need to promote the benefits of using such MIMs; we published a study [Sahar et al., 2023] which emphasises the importance of using MIMs as well as the reasons why the MIMs are not being followed by everyone in the research field.

7.3.2 Integrating the tolerogenic dendritic cells data into a comprehensive and targeted manner

We explained several benefits of efficient data representation and how it has transformed other fields such as drug repurposing or Covid-19 in Chapter 3. We also identified unique issues which hinder data and knowledge representation for subfields in the biology field. We researched different kinds of data representation methods and tools used for different purposes. Finally, we identified that due to the unique nature of the subfields such as

tolDC therapies, no single method can serve the purpose of comprehensive data representation. In Chapter 5, we employed a framework for a targeted and comprehensive data integration in the form of a connected knowledge graph called tolKG.

We made the decision to use Neo4j as the primary platform for integration and data storage in our subsequent work. We opted for a standard graph database instead of a triplestore or RDF for dataset creation due to the anticipated size of the datasets, their specifications and the need for intricate graph-theory-based queries. Graph databases are tailored to support property graphs, where properties can be linked to entities, their relationships, or both. While some triplestores have recently incorporated this feature, it was not part of their original design. Graph databases fundamentally rely on graph theory principles. Given the emphasis on subgraph pattern matching in our mining techniques, a graph database was deemed the best fit.

Due to the lack of data availability in the tolDC therapies field, we designed the data integration framework in a way that it is targeted as well as comprehensive. This needed the integration of different kinds of data sources; we divided these sources into two categories which are unstructured and structured. We used text-mining techniques for the data extraction from the research papers. There were mainly two focuses of the text mining techniques. Primarily the focus was on extracting the biological entities from the targeted set of research papers such as genes, proteins, drugs and chemicals etc. As the project progressed, we identified that it was crucial to extract other non-biological entities also from the text such as author information or the category of the research article. As a result, we employed a combination of strategies to achieve the two goals from the unstructured data set. As a result of these efforts, we could extract 120k entities including genes, proteins, drugs, chemicals etc., and around 1 million relationships between entities.

We developed this framework for the efficient query of tolDC-related data into a knowledge format with the underlying storage and query system. tolKG is the first comprehensive data integration platform for tolDC. Furthermore, it provides a knowledge representation framework which is applicable to other subfields.

7.3.3 Understanding the heterogeneity in the tolDC field

We already discussed in Section 3.2.4 how the tolDC field experiences data heterogeneity due to many different experimental protocols, standardisation issues and analytical methods. However, the extent of this variation is unknown. We used tolKG to explore how protocols differ in terms of reagents and procedures as well as identify common practices across studies. This exploration aims to understand and document the range of experimental designs and cell preparation techniques, highlighting prevalent methodologies that could standardise protocols and advance immunotherapeutic interventions.

We utilised the MITAP guidelines to determine which aspects of heterogeneity in tolDC research can be analyzed through the literature. Starting with the inducing agents for the tolDC, we found rapamycin as the most frequently used inducing agent for tolDC. Although its prevalence is comparable to that of dexamethasone. This might stem from the fact that both vitD3-tolDC and dexa-tolDC often produce cells with similar characteristics, such as a semi-mature phenotype, increased IL-10 secretion and reduced priming of allogeneic T cell proliferation. Researchers might choose between these two based on availability or specific experimental needs, leading to an almost equal combined usage of dexamethasone and vitamin D3 compared to rapamycin, especially in methods-focused studies.

Next, for the growth media, we found that basic media are used more frequently than specialised ones like AIM-V and CellGenix GMP. The exact reason is unknown but one possible reason can be that the basic media can be customised according to the experiments with additional supplements so the researcher prefers to have flexibility. For the maturation process, we observed a trend that all three inducing agents are matured with Cytokine Cocktail or LPS, while the maturation with MPLA is missing. Despite similar outcomes, there is no standardised protocol. There is a need for more standardised approaches to enhance reproducibility and efficiency in the development of tolDC therapies.

Furthermore, we wanted to see if this variation in experiments leads to significantly different pathways so that different types of protocols can be used for different categories of diseases. Our analysis revealed that the disease pathways for dexa-tolDC and vitD3-tolDC are similar while the rapa-tolDC has a different set of pathways. These findings align with the fact that vit D3 and Dexa are similar in properties as inducing agents. We found that some disease categories that are present in the rapa tolDCs are missing from

the other two categories of tolDCs. However, upon further analysis, indirect associations between these diseases and the pathways in dexa-tolDC and vitD3-tolDC were identified. This suggests that certain tolDC types might be more suited for specific diseases, though experimental validation is required to confirm these observations.

7.4 Limitation of the approaches described

Throughout this thesis, our goal was to lay a base for implementing data science methods in the tolDC field. We have tried to achieve it by working on data standardisation and targeted data integration for tolDCs. The methodologies applied here are simple and proven to be efficient. However, the approaches utilised in this thesis also have some limitations. In this section, we identify these limitations and explain them in terms of applicability.

7.4.1 Data selection

We had to carefully select the datasets used at various points in this thesis. In terms of the research paper selection, we utilised PMC and keywords to extract the targeted research papers. While PMC houses a vast collection of articles, it is essential to note that its coverage is not exhaustive. The repository primarily consists of articles that are open access or those for which the authors or publishers have secured permission to be archived in PMC. This means that articles behind paywalls or those that are not open access might not be found in PMC; so we may have missed some of the papers which are not open access.

7.4.2 Validation

We also incorporated the data from other resources such as IntAct or DGidb. Although these databases are well-regarded in the biomedical field, there is always a possibility of errors. Furthermore, in Section 5.5.2, we created edges between the entities based on the external data. We randomly checked a small set of edges, but there is a chance that the framework might have missed some of the edges due to different names.

As mentioned, this is the first approach to performing targeted and comprehensive data integration for the tolDC field, which means there is no previous gold standard

available to test this. This lack of data availability led us to take the help of different kinds of tolDC-related queries. Such queries, as seen in Chapter 6, prove the practicality of tolKG. We also discussed the results with the clinical experts to check if tolKG is accurate. tolKG could serve as a pivotal reference for such upcoming projects in the tolDC field. A unified set of indications will facilitate comparisons across different platforms and ensure validation.

7.5 Future work

This project has carried out an extensive analysis of data standardisation and produced a framework for focused data integration with limited data availability in a specialised field of tolDC therapies. However, there is scope for future improvements. A non-exhaustive discussion of a number of interesting future directions is provided in the following section.

7.5.1 Extending the tolKG

As a Marie Sklodowska-Curie Actions (MSCA) fellow, I was part of a team with 14 researchers, 12 of whom were biomedical researchers. Our original plan was to integrate data from these experts into our data warehouse. However, the pandemic significantly hindered their contributions, as most were unable to access their labs or conduct experiments for a long period. The inclusion of data from them into the tolKG could offer valuable input to the field.

7.5.2 Strategies to update the tolKG

To ensure tolKG remains a valuable and up-to-date resource, it must be continuously expanded and refined to reflect the latest advancements in tolDC research. One key approach is the implementation of automated literature mining pipelines that regularly scan sources such as PubMed for newly published tolDC-related studies. Machine learning-based entity recognition can be used to extract relevant information, allowing seamless integration into tolKG. Additionally, maintaining connections with structured biomedical databases is essential. Establishing real-time or scheduled API integrations with resources like DisGeNET, Reactome and DGIdb will enable the continuous refresh of gene-disease, pathway and drug interaction data, while incorporating patient-derived datasets from

publicly available immunology repositories will further enhance its depth and clinical relevance.

Encouraging community contributions and curation is another important strategy. A user-friendly submission portal can be developed to allow researchers to contribute new tolDC datasets, experimental protocols and findings, while a version control system can track changes and ensure expert validation of new entries. Furthermore, fostering collaborations with research consortia, immunology networks and industry partners will strengthen tolKG's role in the field. Engaging with pharmaceutical companies and electronic lab notebook (ELN) providers can facilitate efficient data-sharing and integration, ensuring that tolKG continues to serve as a dynamic, comprehensive and high-quality knowledge resource for the tolDC research community.

7.5.3 Extending the application of tolKG

If the lab data from researchers were incorporated into the tolKG, we could develop predictive models using machine learning to forecast the outcomes of tolDC therapies in different conditions. This could be particularly useful in customising tolDC protocols and in understanding the factors that lead to desired tolDC outcomes.

7.5.4 Promoting tolKG in the research community

To ensure tolKG has a lasting impact, it must be actively promoted within the relevant research communities. One key strategy is publishing high-impact research papers and review articles that showcase tolKG's applications in immunotherapy and highlight its role in data integration for tolDC therapies. This will help establish tolKG as a recognised resource in the field. Additionally, presenting at major immunology, bioinformatics and systems biology conferences will increase visibility, while organising training workshops will enable researchers to learn how to use tolKG effectively for hypothesis generation and data analysis.

Developing an interactive web platform is another essential step in promoting tolKG. A user-friendly interface where researchers can easily query the knowledge graph, along with downloadable datasets and APIs for seamless integration into bioinformatics pipelines, will encourage broader adoption. Engaging with journals and funding bodies is equally important. Advocacy efforts can focus on encouraging journals to recommend or require

the use of standardised knowledge graphs like tolKG in immunology research, ensuring greater recognition and uptake. Additionally, collaborations with funding agencies can help support projects that contribute new data, further expanding tolKG's relevance and utility. By implementing these strategies, tolKG can evolve into a widely used and continuously updated resource that advances discoveries in tolDC immunotherapy.

7.5.5 Experimental validation of the findings

The primary aim of these findings is to progress promising candidates from computational findings to clinical trials, ultimately benefiting patients. Yet, tolDC therapies are complex and many projects halt at the *in vitro* stage. Validating these findings requires *in vitro* and *in vivo* models like targeted cell assays and mouse experiments. Beyond choosing the right model, selecting suitable candidates for validation is vital. Some solutions might be dismissed by experts due to issues like toxicity, cost or low bioavailability. The financial aspect of clinical trials is also a concern, especially for off-patent drugs. After addressing these considerations, establishing collaborations becomes essential to validate the most promising findings from this project.

7.6 Broader context/Implications

We found that encouraging data sharing in new fields like tolDC therapy is challenging. In addition, it is difficult to enable biologists to share information about their experiments. They know that without the essential information about the experiment, their experiments cannot be repeated but still they do not share the data. Tools like MITAP are designed to facilitate biologists in data sharing but we found from the MITAP analysis that researchers are not using such tools. Even when they use it, they do not use it properly as shown in the MITAP analysis. One of the reasons can be the added burden or the extra training that they require to share their data properly. So they hesitate to share the data at all. This can perhaps be resolved by collaborating with computational or bio-informatics experts in every group.

Developing tolDCs is an expensive, time-consuming and complex process which produces a valuable set of data. The researchers should be motivated to make this valuable

data Findable, Accessible, Interoperable and Reusable (FAIR) to maximize reproducibility and utility as research, and potentially a treatment option. However, data quality and relevancy are also issues in such fields. Because the fields are evolving rapidly, the data becomes irrelevant or out of date. That is why fostering better standards for data quality and interoperability, might often be more beneficial than developing new tools from scratch. This approach can enhance the practical value of biological research tools, making them more effective in aiding the complex analyses required in modern biology.

We also understand that biology is inherently messy, complex and not easily simplified [FarberCancerInstituteinBoston, 2005]. Bioinformatics is a tool that works within this nuanced biological framework and can provide valuable insights. Conversely, alongside the issue of data scarcity, there is the problem of having an abundance of irrelevant data. In some research projects, the bioinformatics aspect is often limited to creating new databases and generating meaningful outcomes from huge datasets and the outcome is not always new. This data is often organized into visually appealing big databases under the assumption that it might prove useful eventually. Generating tolDC cells produces a lot of data but is not always good quality. As a result of this, bioinformatics can create impressively looking visualisations from large sequences which do not provide any new scientific insights, rendering the work ultimately unproductive. This again emphasises the fact that there should be more awareness among researchers for using data standards so that relevant and quality data is shared [Walzer et al., 2013].

Overall, all biological fields go through a shift from a lack of data to a lot of data. The tolDC field is still to reach its peak. Based on the current trends and research focuses, we can predict that the tolDC field will go through significant advancements and broader applications in the coming years. It is also presumable that in the next years, tolDCs will be generated by robots and tolDC therapy might become integrated into personalized medicine approaches where treatments are tailored based on individual immunological profiles. As a result, we will move on from the problem of lack of data in the field to an abundance of data. The research carried out in this thesis will remain relevant as the relevancy of the data that would be included in the research database will still be crucial. Furthermore, the data standards outlined in Chapter 4 will become even more essential as the field evolves.

More importantly, the aim here was to identify and connect the available data on tolDCs as effectively as possible. Based on the database characteristics discussed in

Section 2.2, tolKG can function as a specialised biological database. Although the data in tolKG is not vast due to the niche nature of the field, it is crucial for answering specific questions about tolDCs that would be otherwise unaddressable. The research conducted in this thesis can inform future fields on mitigating data unavailability issues by linking secondary information sources with primary resources, which ultimately enhances the utility of the available data.

7.6.1 Covid-19 pandemic

I began my work on a Marie Skłodowska-Curie Actions (MSCA) project under Horizon 2020 in September 2020, coinciding with the onset of the COVID-19 pandemic. As an MSCA fellow, I collaborated with 14 fellow researchers, 12 of whom are clinical researchers. Our initial plan involved integrating data from these 12 experts into our data warehouse. However, the pandemic severely disrupted their contributions. Most could not access their labs or conduct experiments for an extended period, especially given the time-sensitive nature of generating tolDCs. To this day, we have not received data from any of them.

To adapt, we pivoted towards public data repositories, but the novelty of the field meant limited available data. My PhD journey had two scheduled secondments; only one transpired and that too was virtual. Additionally, the shift of academic conferences to virtual platforms, or their outright cancellation, hindered my chances to showcase my work and connect with peers.

On the positive side, the unprecedented challenges posed by the COVID-19 pandemic underscored the pivotal role of data in guiding both decision-making and research efforts. As researchers grappled with the urgency to understand the novel virus and its ramifications, many turned to repurposing the existing datasets. This encompassed revisiting data from prior epidemics, health records and even seemingly unrelated datasets for potential socio-economic or environmental insights relevant to the spread of the pandemic and impact [Zhang et al., 2022]. In parallel, the search for therapeutic interventions witnessed a surge in drug repurposing efforts, with databases of drug interactions and side effects serving as crucial resources [Mallhi et al., 2021]. Data representation tools, like the dashboard by Johns Hopkins University, became essential for global comprehension [Dong et al., 2020]. The role of semantic data grew, with technologies like ontologies aiding in

merging varied datasets [Wang and He, 2021]. The crisis even spurred the creation of dedicated COVID-19 ontologies to standardise data terminology and relationships [Gretzel et al., 2020].

Due to the pandemic, we transitioned to virtual meetings with the clinical researchers in our group. This eventually turned out to be beneficial in the sense that we were able to engage regularly with more researchers. In addition, we also foresee the long-term effects of COVID-19 on the social aspects of how research is carried out. The pandemic has accelerated the adoption of virtual communication and collaboration tools. This shift is likely to persist, potentially increasing inclusivity by enabling more global collaborations that are less dependent on physical location and travel. COVID-19 also highlighted the importance of rapid research and flexibility in research priorities. This might encourage more adaptive research frameworks that can quickly shift focus in response to global health emergencies. In addition, the pandemic has highlighted the value of open science and rapid data sharing, as seen with the sharing of genomic data and real-time research findings. This trend towards greater openness, if continued, could foster a more collaborative and transparent approach to science.

7.6.2 Impact of advances in knowledge Ggaphs and AI

Since the start of this project, the KG and AI landscape has evolved significantly, particularly with the emergence of large language models (LLMs). These models have transformed knowledge extraction and structuring, offering new possibilities for automating KG construction. If this project were starting now, LLMs would play a greater role in entity and relation extraction, context-aware disambiguation and scalability. While traditional natural language processing methods were effective, LLMs provide improved accuracy and adaptability, though challenges such as hallucination and explainability would need careful management.

Given these advancements, if I were to begin this project today, I would integrate LLM-powered extraction from the outset, adopt hybrid AI approaches combining symbolic reasoning with deep learning and design the KG to evolve dynamically rather than remain static. Additionally, greater emphasis would be placed on explainability and validation to ensure reliability. While this thesis reflects the best practices available at the time, the rapid progress in AI now favours more automated, adaptable, and scalable approaches to

KG development.

7.7 Conclusion

The methodology and tool developed during the course of this project can facilitate standardisation and data integration of existing tolDC data. The work described in this thesis takes into account various issues that are present in the relatively new fields. The work done on the analysis of MITAP and development of tolKG fulfils the aims and research objectives set at the start of this thesis (see Section 1.3). Besides the main contributions of this project, the approaches used during tolKG can also be used to build a sufficient corpus for conducting comprehensive literature reviews. Furthermore, tolkG can be used for applying machine learning and AI. Breakthroughs in technology, especially in areas like genomics, bioinformatics and molecular imaging, provide tools that allow for deeper and more detailed studies than previously possible. This technological progress enables researchers to explore new dimensions of biology, leading to the creation of specialised subfields. The approaches and ideas explored during this project can be applied to such new fields also.

This project is a multi-disciplinary project in a new field of biology, which means we have navigated numerous challenges, from a scarcity of relevant data to a general lack of awareness about data sharing within the field. The insights gained from this project are not only practically beneficial but also instrumental in promoting data-sharing awareness among researchers. We are developing new methods to address previously untreatable deadly diseases, yet the efforts to enhance the reusability of scientific research remain modest. The transition from theoretical data-sharing frameworks like the five stars of data and FAIR data principles to practical application needs to expand beyond major research groups to include smaller lab groups, where many new fields begin. This broader application is essential for fostering innovation and collaboration in burgeoning areas of research. The work presented here will also benefit greatly from the improved use of data standards and FAIR data sharing in the tolDC field.

Finally, we hope that the findings of this project will be beneficial to the broader community and ultimately contribute to enhancing the well-being of individuals affected by immune diseases.

Bibliography

[imm, 2020] (2020). ImmGen at 15. *Nature Immunology*, 21(7):700–703. 39

[Athar et al., 2019] Athar, A., Füllgrabe, A., George, N., Iqbal, H., Huerta, L., Ali, A., Snow, C., Fonseca, N. A., Petryszak, R., Papatheodorou, I., et al. (2019). ArrayExpress update–from bulk to single-cell expression data. *Nucleic acids research*, 47(D1):D711–D715. 39

[Banchereau and Steinman, 1998] Banchereau, J. and Steinman, R. M. (1998). Dendritic cells and the control of immunity. *Nature*, 392(6673):245–252. 33

[Banerjee et al., 2020] Banerjee, A., Chakraborty, C., Kumar, A., and Biswas, D. (2020). Emerging trends in IoT and big data analytics for biomedical and health care technologies. *Handbook of data science approaches for biomedical engineering*, pages 121–152. 21

[Barrett et al., 2012] Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., et al. (2012). NCBI GEO: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1):D991–D995. 39

[Begley and Ellis, 2012] Begley, C. and Ellis, L. (2012). Drug development: Raise standards for preclinical cancer research. Nature.[Online]. 483 (7391). 6

[Bell et al., 2017] Bell, G., Anderson, A., Diboll, J., Reece, R., Eltherington, O., Harry, R., Fouweather, T., MacDonald, C., Chadwick, T., McColl, E., et al. (2017). Autologous tolerogenic dendritic cells for rheumatoid and inflammatory arthritis. *Annals of the rheumatic diseases*, 76(1):227–234. 33

[Beltagy et al., 2019] Beltagy, I., Lo, K., and Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*. 78

[Bendels et al., 2018] Bendels, M. H., Müller, R., Brueggmann, D., and Groneberg, D. A. (2018). Gender disparities in high-quality research revealed by Nature Index journals. *PloS one*, 13(1):e0189136. 127

[Benham et al., 2015] Benham, H., Nel, H. J., Law, S. C., Mehdi, A. M., Street, S., Ramnoruth, N., Pahau, H., Lee, B. T., Ng, J., G. Brunck, M. E., et al. (2015). Citrullinated peptide dendritic cell immunotherapy in HLA risk genotype–positive rheumatoid arthritis patients. *Science translational medicine*, 7(290):290ra87–290ra87. 33

[Berman et al., 2000] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242. 39

[Berners-Lee et al., 2001] Berners-Lee, T., Hendler, J., and Lassila, O. (2001). A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Scientific american*, 284(5):34–43. 24

[Bhattacharya et al., 2018] Bhattacharya, S., Dunn, P., Thomas, C. G., Smith, B., Schaefer, H., Chen, J., Hu, Z., Zalocusky, K. A., Shankar, R. D., Shen-Orr, S. S., et al. (2018). ImmPort, toward repurposing of open access immunological assay data for translational and clinical research. *Scientific data*, 5(1):1–9. 21, 39

[Bodenreider, 2006] Bodenreider, O. (2006). Lexical, terminological and ontological resources for biological text mining. *Text mining for biology and biomedicine*, pages 43–66. 45

[Bona et al., 2019] Bona, J. P., Prior, F. W., Zozus, M. N., and Brochhausen, M. (2019). Enhancing clinical data and clinical research data with biomedical ontologies-insights from the knowledge representation perspective. *Yearbook of medical informatics*, 28(01):140–151. 24

[Boulton, 2016] Boulton, G. (2016). International accord on open data. *Nature*, 530(7590):281–281. 37

[Brazma et al., 2001] Brazma, A., Hingamp, P., Quackenbush, J., Sherlock, G., Spellman, P., Stoeckert, C., Aach, J., Ansorge, W., Ball, C. A., Causton, H. C., et al. (2001). Minimum information about a microarray experiment (MIAME)—toward standards for microarray data. *Nature genetics*, 29(4):365–371. 46, 69

[Brazma et al., 2003] Brazma, A., Parkinson, H., Sarkans, U., Shojatalab, M., Vilo, J., Abeygunawardena, N., Holloway, E., Kapushesky, M., Kemmeren, P., Lara, G. G., et al. (2003). ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic acids research*, 31(1):68–71. 46

[Breuer et al., 2013] Breuer, K., Foroushani, A. K., Laird, M. R., Chen, C., Sribnaia, A., Lo, R., Winsor, G. L., Hancock, R. E., Brinkman, F. S., and Lynn, D. J. (2013). InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic acids research*, 41(D1):D1228–D1233. 21

[Bukhari et al., 2019] Bukhari, S. A. C., Mandell, J., Kleinstein, S. H., and Cheung, K.-H. (2019). A linked data graph approach to integration of immunological data. In *Proceedings. IEEE International Conference on Bioinformatics and Biomedicine*, volume 2019, page 1742. NIH Public Access. 21, 39

[Canning et al., 2001] Canning, M., Grotenhuis, K., Ruwhof, C., Drexhage, H., and de Wit, H. (2001). 1-alpha, 25-Dihydroxyvitamin D3 (1, 25 (OH)(2) D (3)) hampers the maturation of fully active immature dendritic cells from monocytes. *European Journal of Endocrinology*. 33

[Chen et al., 2021] Chen, C., Ross, K. E., Gavali, S., Cowart, J. E., and Wu, C. H. (2021). COVID-19 Knowledge Graph from semantic integration of biomedical literature and databases. *Bioinformatics*, 37(23):4597–4598. 22

[Chen and Sharp, 2004] Chen, H. and Sharp, B. M. (2004). Content-rich biological network constructed by mining PubMed abstracts. *BMC bioinformatics*, 5:1–13. 45

[Conrad et al., 2023] Conrad, N., Misra, S., Verbakel, J. Y., Verbeke, G., Molenberghs, G., Taylor, P. N., Mason, J., Sattar, N., McMurray, J. J., McInnes, I. B., et al. (2023). Incidence, prevalence, and co-occurrence of autoimmune disorders over time and by age, sex, and socioeconomic status: a population-based cohort study of 22 million individuals in the UK. *The Lancet*, 401(10391):1878–1890. 6

[Cotto et al., 2018] Cotto, K. C., Wagner, A. H., Feng, Y.-Y., Kiwala, S., Coffman, A. C., Spies, G., Wollam, A., Spies, N. C., Griffith, O. L., and Griffith, M. (2018). DGIdb 3.0: a redesign and expansion of the drug–gene interaction database. *Nucleic acids research*, 46(D1):D1068–D1073. 80, 86

[Dall'Olio et al., 2010] Dall'Olio, G. M., Bertranpetit, J., and Laayouni, H. (2010). The annotation and the usage of scientific databases could be improved with public issue tracker software. *Database*, 2010. 39

[Darmoni et al., 2012] Darmoni, S. J., Soualmia, L. F., Letord, C., Jaulent, M.-C., Griffon, N., Thirion, B., and Névéol, A. (2012). Improving information retrieval using Medical Subject Headings Concepts: a test case on rare and chronic diseases. *Journal of the Medical Library Association: JMLA*, 100(3):176. 82

[Deng et al., 2022] Deng, N., Wu, C., Yaseen, A., and Wu, H. (2022). ImmuneData: an integrated data discovery system for immunology data repositories. *Database*, 2022. 39

[Domingo-Fernández et al., 2021] Domingo-Fernández, D., Baksi, S., Schultz, B., Gadiya, Y., Karki, R., Raschka, T., Ebeling, C., Hofmann-Apitius, M., and Kodamullil, A. T. (2021). COVID-19 Knowledge Graph: a computable, multi-modal, cause-and-effect knowledge model of COVID-19 pathophysiology. *Bioinformatics*, 37(9):1332–1334. 22

[Dong et al., 2020] Dong, E., Du, H., and Gardner, L. (2020). An interactive web-based dashboard to track COVID-19 in real time. *The Lancet infectious diseases*, 20(5):533–534. 142

[Edgar et al., 2002] Edgar, R., Domrachev, M., and Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic acids research*, 30(1):207–210. 46

[Ehrlinger and Wöß, 2016] Ehrlinger, L. and Wöß, W. (2016). Towards a definition of knowledge graphs. *SEMANTiCS (Posters, Demos, SuCCESS)*, 48(1-4):2. 25

[Fabregat et al., 2018] Fabregat, A., Korninger, F., Viteri, G., Sidiropoulos, K., Marin-Garcia, P., Ping, P., Wu, G., Stein, L., D'Eustachio, P., and Hermjakob, H. (2018). Reactome graph database: Efficient access to complex pathway data. *PLoS computational biology*, 14(1):e1005968. 86, 92

[Falcón-Beas et al., 2019] Falcón-Beas, C., Tittarelli, A., Mora-Bau, G., Tempio, F., Pérez, C., Hevia, D., Behrens, C., Flores, I., Falcón-Beas, F., Garrido, P., et al. (2019). Dexamethasone turns tumor antigen-presenting cells into tolerogenic dendritic cells with T cell inhibitory functions. *Immunobiology*, 224(5):697–705. 33

[FarberCancerInstituteinBoston, 2005] FarberCancerInstituteinBoston, M. (2005). Playingdirty. 141

[Field et al., 2008] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., et al. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature biotechnology*, 26(5):541–547. 46, 69

[Fuchs et al., 2018] Fuchs, A., Gliwiński, M., Grageda, N., Spiering, R., Abbas, A. K., Appel, S., Bacchetta, R., Battaglia, M., Berglund, D., Blazar, B., et al. (2018). Minimum information about T regulatory cells: a step toward reproducibility and standardization. *Frontiers in Immunology*, 8:1844. 70

[Fucikova et al., 2019] Fucikova, J., Palova-Jelinkova, L., Bartunkova, J., and Spisek, R. (2019). Induction of tolerance and immunity by dendritic cells: mechanisms and clinical applications. *Frontiers in immunology*, 10:2393. 54

[Gajzler, 2016] Gajzler, M. (2016). Usefulness of mining methods in knowledge source analysis in the construction industry. *Archives of Civil Engineering*, 62(1):127–142. xi, 3, 5

[Gan et al., 2019] Gan, J., Cai, Q., Galer, P., Ma, D., Chen, X., Huang, J., Bao, S., and Luo, R. (2019). Mapping the knowledge structure and trends of epilepsy genetics over the past decade: A co-word analysis based on medical subject headings terms. *Medicine*, 98(32). 82

[Garcia, 2018] Garcia, P. (2018). Are you# AutoimmuneAware? Report for parliamentarians into autoimmune conditions. *JDRF*. 6

[Giannoukakis et al., 2011] Giannoukakis, N., Phillips, B., Finegold, D., Harnaha, J., and Trucco, M. (2011). Phase I (safety) study of autologous tolerogenic dendritic cells in type 1 diabetic patients. *Diabetes care*, 34(9):2026–2032. 33

[Gosselin, 2021] Gosselin, R.-D. (2021). Insufficient transparency of statistical reporting in preclinical research: a scoping review. *Scientific Reports*, 11(1):1–8. 41

[Goulart et al., 2011] Goulart, R. R. V., Strube de Lima, V. L., and Xavier, C. C. (2011). A systematic review of named entity recognition in biomedical texts. *Journal of the Brazilian Computer Society*, 17(2):103–116. 83

[Gretzel et al., 2020] Gretzel, U., Fuchs, M., Baggio, R., Hoepken, W., Law, R., Neidhardt, J., Pesonen, J., Zanker, M., and Xiang, Z. (2020). e-Tourism beyond COVID-19: a call for transformative research. *Information Technology & Tourism*, 22:187–203. 143

[Grigorian-Shamagian et al., 2021] Grigorian-Shamagian, L., Sanz-Ruiz, R., Climent, A., Badimon, L., Barile, L., Bolli, R., Chamuleau, S., Grobbee, D. E., Janssens, S., Kastrup, J., et al. (2021). Insights into therapeutic products, preclinical research models, and clinical trials in cardiac regenerative and reparative medicine: where are we now and the way ahead. Current opinion paper of the ESC Working Group on Cardiovascular Regenerative and Reparative Medicine. *Cardiovascular research*, 117(6):1428–1433. 41

[Hilkens and Isaacs, 2013] Hilkens, C. and Isaacs, J. (2013). Tolerogenic dendritic cell therapy for rheumatoid arthritis: where are we now? *Clinical & Experimental Immunology*, 172(2):148–157. 41

[Hippen and Greene, 2021] Hippen, A. A. and Greene, C. S. (2021). Expanding and remixing the metadata landscape. *Trends in cancer*, 7(4):276–278. 47

[Hirschman et al., 2005] Hirschman, L., Yeh, A., Blaschke, C., and Valencia, A. (2005). Overview of BioCreAtIvE: critical assessment of information extraction for biology. 45

[Hunter and Cohen, 2006] Hunter, L. and Cohen, K. B. (2006). Biomedical language processing: what's beyond PubMed? *Molecular cell*, 21(5):589–594. 45

[Ioannidis, 2005] Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8):e124. 37

[Janetzki et al., 2009] Janetzki, S., Britten, C. M., Kalos, M., Levitsky, H. I., Maecker, H. T., Melief, C. J., Old, L. J., Romero, P., Hoos, A., and Davis, M. M. (2009). "MIATA"—minimal information about T cell assays. *Immunity*, 31(4):527–528. 70

[Jassal et al., 2020] Jassal, B., Matthews, L., Viteri, G., Gong, C., Lorente, P., Fabregat, A., Sidiropoulos, K., Cook, J., Gillespie, M., Haw, R., et al. (2020). The reactome pathway knowledgebase. *Nucleic acids research*, 48(D1):D498–D503. 80

[Kahn Jr, 2022] Kahn Jr, C. E. (2022). Analysis of Causal Relationships in Integrated Ontologies of Diseases, Phenotypes, and Radiological Diagnosis. In *MEDINFO 2021: One World, One Health—Global Partnership for Digital Innovation: Proceedings of the 18th World Congress on Medical and Health Informatics*, volume 290, page 258. IOS Press. 25

[Kerrien et al., 2012] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2012). The IntAct molecular interaction database in 2012. *Nucleic acids research*, 40(D1):D841–D846. 86

[Kim et al., 2017] Kim, J., Kim, J.-j., and Lee, H. (2017). An analysis of disease-gene relationship from Medline abstracts by DigSee. *Scientific reports*, 7(1):1–13. 104

[Kostoff, 2014] Kostoff, R. N. (2014). Literature-related discovery: common factors for Parkinson's Disease and Crohn's Disease. *Scientometrics*, 100(3):623–657. 104

[Kurochkina et al., 2018] Kurochkina, Y., Tikhonova, M., Tyrinova, T., Leplina, O., Sizikov, A., Sulutian, A., Chumasova, O., Ostanin, A., and Chernykh, E. (2018). SAT0212 The safety and tolerability of intra-articular injection of tolerogenic dendritic cells in patients with rheumatoid arthritis: the preliminary results. 33

[Le and Mikolov, 2014] Le, Q. and Mikolov, T. (2014). Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR. 89

[Lee et al., 2020] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., and Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240. 78

[Lerner et al., 2015] Lerner, A., Jeremias, P., and Matthias, T. (2015). The world incidence and prevalence of autoimmune diseases is increasing. *Int J Celiac Dis*, 3(4):151–5. 6

[Li et al., 2023] Li, S., Wong, K. W., Zhu, D., and Fung, C. C. (2023). Drug-CoV: a drug-origin knowledge graph discovering drug repurposing targeting COVID-19. *Knowledge and Information Systems*, 65(12):5289–5308. 22

[Li, 2016] Li, X. (2016). A meaning-oriented approach to semantic data modeling. *arXiv preprint arXiv:1609.03346*. 24

[Linnaeus, 1758] Linnaeus, C. (1758). *Systema naturae per regna tria naturae, secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.* Laurentius Salvius. 15

[Lord et al., 2016] Lord, P., Spiering, R., Aguillon, J. C., Anderson, A. E., Appel, S., Benitez-Ribas, D., Ten Brinke, A., Broere, F., Cools, N., Cuturi, M. C., et al. (2016). Minimum information about tolerogenic antigen-presenting cells (MITAP): a first step towards reproducibility and standardisation of cellular therapies. *PeerJ*, 4:e2300. 37, 46, 53, 73

[Maecker et al., 2012] Maecker, H. T., McCoy, J. P., and Nussenblatt, R. (2012). Standardizing immunophenotyping for the human immunology project. *Nature Reviews Immunology*, 12(3):191–200. 43

[Mallhi et al., 2021] Mallhi, T. H., Khan, Y. H., Alotaibi, N. H., Alzarea, A. I., Alanazi, A. S., Qasim, S., Iqbal, M. S., and Tanveer, N. (2021). Drug repurposing for COVID-19: a potential threat of self-medication and controlling measures. *Postgraduate medical journal*, 97(1153):742–743. 142

[Marin-Gallen et al., 2010] Marin-Gallen, S., Clemente-Casares, X., Planas, R., Pujol-Autonell, I., Carrascal, J., Carrillo, J., Ampudia, R., Verdaguer, J., Pujol-Borrell, R., Borras, F., et al. (2010). Dendritic cells pulsed with antigen-specific apoptotic bodies prevent experimental type 1 diabetes. *Clinical & Experimental Immunology*, 160(2):207–214. 33

[McGuinness, 2007] McGuinness, D. L. (2007). Owl web ontology language overview w3c recommendation 10 february 2004. *http://www. w3. org/TR/2004/REC-owl-features-20040210/*. 24

[Merabti et al., 2012] Merabti, T., Soualmia, L. F., Grosjean, J., Joubert, M., and Darmoni, S. J. (2012). Aligning biomedical terminologies in French: towards semantic interoperability in medical applications. *Medical Informatics*, pages 41–68. 24

[Miller, 2013] Miller, J. J. (2013). Graph database applications and concepts with Neo4j. In *Proceedings of the southern association for information systems conference, Atlanta, GA, USA*, volume 2324. 92

[Moreau et al., 2012] Moreau, A., Varey, E., Bouchet-Delbos, L., and Cuturi, M.-C. (2012). Cell therapy using tolerogenic dendritic cells in transplantation. *Transplantation research*, 1:1–8. 33

[Morelli and Thomson, 2007] Morelli, A. E. and Thomson, A. W. (2007). Tolerogenic dendritic cells and the quest for transplant tolerance. *Nature Reviews Immunology*, 7(8):610–621. 33

[Mosanya and Isaacs, 2019] Mosanya, C. H. and Isaacs, J. D. (2019). Tolerising cellular therapies: what is their promise for autoimmune disease? *Annals of the Rheumatic Diseases*, 78(3):297–310. 33

[Mullen et al., 2016] Mullen, J., Cockell, S. J., Woollard, P., and Wipat, A. (2016). An integrated data driven approach to drug repositioning using gene-disease associations. *PloS one*, 11(5):e0155811. 22

[Munafò et al., 2017] Munafò, M. R., Nosek, B. A., Bishop, D. V., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., and Ioannidis, J. (2017). A manifesto for reproducible science. *Nature human behaviour*, 1(1):1–9. 6

[Nadeau and Sekine, 2007] Nadeau, D. and Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26. 23

[Nadif and Role, 2021] Nadif, M. and Role, F. (2021). Unsupervised and self-supervised deep learning approaches for biomedical text mining. *Briefings in Bioinformatics*, 22(2):1592–1603. 83

[Naseem et al., 2022] Naseem, U., Dunn, A. G., Khushi, M., and Kim, J. (2022). Benchmarking for biomedical natural language processing tasks with a domain specific ALBERT. *BMC bioinformatics*, 23(1):144. 78

[Navarro-Barriuso et al., 2018] Navarro-Barriuso, J., Mansilla, M. J., Naranjo-Gómez, M., Sánchez-Pla, A., Quirant-Sánchez, B., Teniente-Serra, A., Ramo-Tello, C., and Martínez-Cáceres, E. M. (2018). Comparative transcriptomic profile of tolerogenic dendritic cells differentiated with vitamin D3, dexamethasone and rapamycin. *Scientific reports*, 8(1):1–13. 74, 122

[Niedzwiedz et al., 2019] Niedzwiedz, C. L., Knifton, L., Robb, K. A., Katikireddi, S. V., and Smith, D. J. (2019). Depression and anxiety among people living with and beyond cancer: a growing clinical and research priority. *BMC cancer*, 19(1):1–8. 124

[Olson, 1993] Olson, M. V. (1993). The human genome project. *Proceedings of the National Academy of Sciences*, 90(10):4338–4344. 15

[Orchard et al., 2014] Orchard, S., Ammari, M., Aranda, B., Breuza, L., Briganti, L., Broackes-Carter, F., Campbell, N. H., Chavali, G., Chen, C., Del-Toro, N., et al. (2014). The MIntAct project—IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic acids research*, 42(D1):D358–D363. 79

[Ostaszewski et al., 2020] Ostaszewski, M., Mazein, A., Gillespie, M. E., Kuperstein, I., Niarakis, A., Hermjakob, H., Pico, A. R., Willighagen, E. L., Evelo, C. T., Hasenauer, J., et al. (2020). COVID-19 Disease Map, building a computational repository of SARS-CoV-2 virus-host interaction mechanisms. *Scientific data*, 7(1):136. 22

[Park et al., 2014] Park, Y., Shankar, M., Park, B.-H., and Ghosh, J. (2014). Graph databases for large-scale healthcare systems: A framework for efficient data management and data services. In *2014 IEEE 30th International Conference on Data Engineering Workshops*, pages 12–19. IEEE. 27

[Peng et al., 2020] Peng, Y., Chen, Q., and Lu, Z. (2020). An empirical study of multi-task learning on BERT for biomedical text mining. *arXiv preprint arXiv:2005.02799*. 83

[Piñero et al., 2016] Piñero, J., Bravo, À., Queralt-Rosinach, N., Gutiérrez-Sacristán, A., Deu-Pons, J., Centeno, E., García-García, J., Sanz, F., and Furlong, L. I. (2016). DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, page gkw943. 80, 86

[Popovski et al., 2019] Popovski, G., Kochev, S., Korousic-Seljak, B., and Eftimov, T. (2019). FoodIE: A Rule-based Named-entity Recognition Method for Food Information Extraction. *ICPRAM*, 12:pp–915. 83

[Res, 2012] Res, A. (2012). Jan; 40 (Database issue): D700-5. *PubMed PMID*, 22110037. 39

[Riccaboni and Verginer, 2022] Riccaboni, M. and Verginer, L. (2022). The impact of the COVID-19 pandemic on scientific research in the life sciences. *PLoS One*, 17(2):e0263001. 82

[Robinson et al., 2015] Robinson, I., Webber, J., and Eifrem, E. (2015). *Graph databases: new opportunities for connected data*. " O'Reilly Media, Inc.". 27

[Sahar et al., 2023] Sahar, A., Nicorescu, I., Barran, G., Paterson, M., Hilkens, C. M., and Lord, P. (2023). Tolerogenic dendritic cell reporting: Has a minimum information model made a difference? *PeerJ*, 11:e15352. iv, 11, 53, 134

[Sang et al., 2018] Sang, S., Yang, Z., Wang, L., Liu, X., Lin, H., and Wang, J. (2018). SemaTyP: a knowledge graph based literature mining method for drug discovery. *BMC bioinformatics*, 19(1):1–11. 104

[Savova et al., 2010] Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., and Chute, C. G. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513. 23

[Sayers et al., 2010] Sayers, E. W., Barrett, T., Benson, D. A., Bolton, E., Bryant, S. H., Canese, K., Chetvernin, V., Church, D. M., DiCuccio, M., Federhen, S., et al. (2010). Database resources of the national center for biotechnology information. *Nucleic acids research*, 39(suppl_1):D38–D51. 82

[Schultze and Aschenbrenner, 2021] Schultze, J. L. and Aschenbrenner, A. C. (2021). COVID-19 and the human innate immune system. *Cell*, 184(7):1671–1692. 43

[Schwartz and Hearst, 2002] Schwartz, A. S. and Hearst, M. A. (2002). A simple algorithm for identifying abbreviation definitions in biomedical text. In *Biocomputing 2003*, pages 451–462. World Scientific. 83

[SciBite, 2023] SciBite (2023). Semantic Analytics: A systematic, data-driven approach to drug repositioning, url: https://scibite.com/resources/drug-repositioning-whitepaper/. 84

[Singhal et al., 2012] Singhal, A. et al. (2012). Introducing the knowledge graph: things, not strings. *Official google blog*, 5(16):3. 25

[Smith et al., 2007] Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, 25(11):1251–1255. 24

[Snickars and Weibull, 1977] Snickars, F. and Weibull, J. W. (1977). A minimum information principle: Theory and practice. *Regional science and urban economics*, 7(1-2):137–168. 46

[Song et al., 2021] Song, B., Li, F., Liu, Y., and Zeng, X. (2021). Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison. *Briefings in Bioinformatics*, 22(6):bbab282. 83

[Song et al., 2015] Song, M., Kim, W. C., Lee, D., Heo, G. E., and Kang, K. Y. (2015). PKDE4J: Entity and relation extraction for public knowledge discovery. *Journal of biomedical informatics*, 57:320–332. 104

[Stenger et al., 2014] Stenger, E. O., Rosborough, B. R., Mathews, L. R., Ma, H., Mapara, M. Y., Thomson, A. W., and Turnquist, H. R. (2014). IL-12hi Rapamycin-Conditioned Dendritic Cells Mediate IFN-γ–Dependent Apoptosis of Alloreactive CD4+ T Cells In Vitro and Reduce Lethal Graft-Versus-Host Disease. *Biology of Blood and Marrow Transplantation*, 20(2):192–201. 33

[Suuring and Moreau, 2021] Suuring, M. and Moreau, A. (2021). Regulatory Macrophages and Tolerogenic Dendritic Cells in Myeloid Regulatory Cell-Based Therapies. *International Journal of Molecular Sciences*, 22(15):7970. 33

[Szklarczyk et al., 2021] Szklarczyk, D., Gable, A. L., Nastou, K. C., Lyon, D., Kirsch, R., Pyysalo, S., Doncheva, N. T., Legeay, M., Fang, T., Bork, P., et al. (2021). The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic acids research*, 49(D1):D605–D612. 104

[Taylor et al., 2008] Taylor, C. F., Field, D., Sansone, S.-A., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P.-A., Bogue, M., Booth, T., et al. (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. *Nature biotechnology*, 26(8):889–896. 54

[Ten Brinke et al., 2015] Ten Brinke, A., Hilkens, C. M., Cools, N., Geissler, E. K., Hutchinson, J. A., Lombardi, G., Lord, P., Sawitzki, B., Trzonkowski, P., Van Ham, S. M., et al. (2015). Clinical use of tolerogenic dendritic cells-harmonization approach in European collaborative effort. *Mediators of inflammation*, 2015. 41

[Toussi and Massari, 2014] Toussi, D. N. and Massari, P. (2014). Immune adjuvant effect of molecularly-defined toll-like receptor ligands. *Vaccines*, 2(2):323–353. 43

[Turnquist et al., 2007] Turnquist, H. R., Raimondi, G., Zahorchak, A. F., Fischer, R. T., Wang, Z., and Thomson, A. W. (2007). Rapamycin-conditioned dendritic cells are poor stimulators of allogeneic CD4+ T cells, but enrich for antigen-specific Foxp3+ T regulatory cells and promote organ transplant tolerance. *The Journal of Immunology*, 178(11):7018–7031. 33

[Vechina et al., 2013] Vechina, A., Arrais, J., and Oliveira, J. L. (2013). Representation of Semantic Networks of Biomedical Terms. In *IWBBIO*, pages 703–710. 24

[Walzer et al., 2013] Walzer, M., Qi, D., Mayer, G., Uszkoreit, J., Eisenacher, M., Sachsenberg, T., Gonzalez-Galarza, F. F., Fan, J., Bessant, C., Deutsch, E. W., et al. (2013). The mzquantml data standard for mass spectrometry–based quantitative studies in proteomics. *Molecular & cellular proteomics*, 12(8):2332–2340. 141

[Wang et al., 2014] Wang, S., Pandis, I., Wu, C., He, S., Johnson, D., Emam, I., Guitton, F., and Guo, Y. (2014). High dimensional biological data retrieval optimization with NoSQL technology. In *BMC genomics*, volume 15, pages 1–8. Springer. 92

[Wang and He, 2021] Wang, Z. and He, Y. (2021). Precision omics data integration and analysis with interoperable ontologies and their application for COVID-19 research. *Briefings in Functional Genomics*, 20(4):235–248. 143

[Warrender and Lord, 2015] Warrender, J. D. and Lord, P. (2015). Scaffolding the Mitochondrial Disease Ontology from extant knowledge sources. *arXiv preprint arXiv:1505.04114*. 105

[Wei et al., 2015] Wei, C.-H., Kao, H.-Y., Lu, Z., et al. (2015). GNormPlus: an integrative approach for tagging genes, gene families, and protein domains. *BioMed research international*, 2015. 84

[Willekens et al., 2019] Willekens, B., Presas-Rodríguez, S., Mansilla, M., Derdelinckx, J., Lee, W.-P., Nijs, G., De Laere, M., Wens, I., Cras, P., Parizel, P., et al. (2019). Tolerogenic dendritic cell-based treatment for multiple sclerosis (MS): a harmonised study protocol for two phase I clinical trials comparing intradermal and intranodal cell administration. *BMJ open*, 9(9):e030309. 33

[Zhang et al., 2022] Zhang, Q., Gao, J., Wu, J. T., Cao, Z., and Dajun Zeng, D. (2022). Data science approaches to confronting the COVID-19 pandemic: a narrative review. *Philosophical Transactions of the Royal Society A*, 380(2214):20210127. 142

[Zubizarreta et al., 2019] Zubizarreta, I., Flórez-Grau, G., Vila, G., Cabezón, R., España, C., Andorra, M., Saiz, A., Llufriu, S., Sepulveda, M., Sola-Valls, N., et al. (2019). Immune tolerance in multiple sclerosis and neuromyelitis optica with peptide-loaded tolerogenic dendritic cells in a phase 1b trial. *Proceedings of the National Academy of Sciences*, 116(17):8463–8470. 33

A

Appendix

FIGURE A.1: Complete heatmap of the MITAP compliant papers and Non MITAP papers Green: category reported in the publication; Yellow: category partially reported in the publication; Red circle: category unreported in the publication.