DISSERTATION | DOCTORAL THESIS

Titel | Title

Toxicity Predictions of Biological Drugs

verfasst von | submitted by

Abir Omran B.Sc. farm. M.

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Doktorin der Naturwissenschaften (Dr.rer.nat.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 796 610 449
Dissertationsgebiet lt. Studienblatt | Field of study as it appears on the student record sheet:	Pharmazie
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Gerhard Ecker

Declaration of Author

I hereby declare that the presented work in this thesis is my original work, which I carried out myself. All sources, references, and literature used during the elaboration of this work are properly cited and listed in complete reference to the due source.

Vienna, May 2025, Abir Omran

Table of Contents

Preface

The work that is presented in this thesis was performed under the supervision of Prof. Gerhard F. Ecker in the Pharmacoinformatics Research Group at University of Vienna and in collaboration with the Preclinical Safety group at Sanofi.

Chapter 1 provides a brief introduction to drug discovery and biological drugs, as well as to the toxicity of biological drugs and in silico methods for assessing their toxicity.

Chapter 2 focuses on some of the methodologies used and applied in this work.

Chapter 3 focuses on the aim of the thesis and the research questions.

Chapter 4 presents the three studies which are the results of the thesis. Study 1 presents the interactive dashboard that was developed for analyzing and visualizing side effect data, whereas Study 2 focuses on the analysis of side effect data for therapeutic monoclonal antibodies. Study 3 demonstrates the employment of T-cell assay data for machine-learning models for immunogenicity prediction.

Chapter 5 discusses the three studies and contains the conclusion and outlook of the thesis.

This PhD project was a part of the MolTag doctoral program (FWF grant W1232) and funded by Sanofi.

Chapter 1: Introduction

Drug discovery

The drug discovery and development pipeline is a long and complex process that can be divided into five stages: discovery and development, preclinical research, clinical research, review and approval, and post-market safety monitoring [1]. Within the discovery and development stage, a disease is investigated and a potential target is identified. This is followed by the drug discovery phase, where molecules are screened for interaction with the intended target and evaluated for their potential biological effects. Once a promising compound has been identified, it is further optimized to enhance efficacy and reduce toxicity. The preclinical stage is essential for assessing activity and toxicity before the drug is tested in humans for the first time. This is done *in vivo*, where an animal model — selected based on its ability to best mimic human responses is used, and *in vitro*, where human cells are employed [2]. Following the preclinical stage is the clinical phase, during which the drug is tested on humans for the first time. It is therefore critical that prior stages provide sufficient data and understanding of the drug to ensure its safety and efficacy before human exposure. If the drug is approved, it enters post-marketing surveillance, where it continues to be monitored in real-word settings. Given the time-consuming and expensive process, leveraging insight from previous drug development efforts is essential to improve the likelihood of success. Additionally, there is a pressing need for methods and tools that enable more efficient testing while simultaneously reducing unnecessary use of animals and human subjects [2].

Biological drugs

The regulatory paths in some stages differ between small molecule drugs and biological drugs [3]. Biological drugs, also referred to as biologics, are molecules derived from living organism or biological process. They include vaccines, peptides, and monoclonal antibodies (mAbs), among others [3]. Biologics are usually therapeutics proteins, with a larger and more complex structure in comparison to small molecules. Figure 1 provides a schematic overview of an antibody, illustrating the general structure. Biologics, such as mAbs, have shown high efficacy and have succeeded in treating diseases that were once considered untreatable [4]. In addition, they are more selective than small molecule drugs and, therefore, cause fewer off-

target side effects. These are some of the reasons why biologics are receiving growing scientific attention [4].

However, although biologics cause fewer off-target side effects, they can still cause severe side effects, which in some cases may lead to life-threatening consequences or even death [5]. As the demand for biological drugs increases and more are being approved, it becomes even more critical to understand their side effect profiles in order to better assess their potential risks.

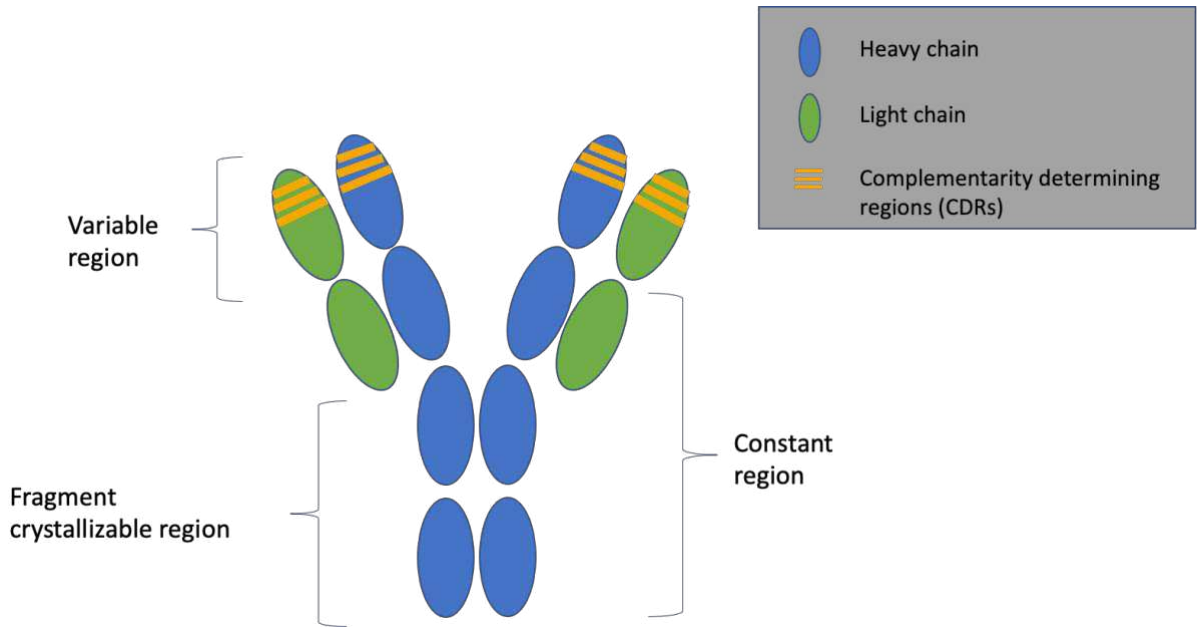


Figure 1: The figure illustrates the general structure of an antibody and the location of key regions, such as the complementarity determining regions (CDRs) which are the binding regions.

Toxicity of biological drugs

As previously mentioned, biological drugs are highly specific and therefore result in less off-target toxicity compared to small molecules. Nonetheless, drugs such as Infliximab and Adalimumab, which are generally not associated with severe side effects, have been reported to cause drug-induced liver injury (DILI) [6]. The exact mechanism is not fully understood

[6]. That said, it has been shown that the majority of affected patients exhibited autoimmune features while experiencing DILI [6].

Biologics are inherently immunogenic, whereas small molecule drugs are generally not. For vaccines, immunogenicity is a desired outcome as their effectiveness depend on eliciting an immune response. In contrast, for other drugs, such as certain therapeutic mAbs, an immune response is considered an undesirable side effect. Therapeutic mAbs can include components derived from different species, and the less human the antibody is, the more likely it is to be immunogenic. There are four main types of mAbs and are presented in Figure 2. The types are: murine antibodies, which are entirely of non-human origin; chimeric antibodies, which have human constant regions and non-human variable regions; humanized antibodies, in which only the complement-determining regions (CDRs), which are the binding regions, are of non-human origin; and fully human antibodies. Immunogenicity in therapeutic mAbs typically leads to the production of anti-drug antibodies (ADAs), which can bind to the mAbs and result in a loss of efficacy. Since most therapeutic mAbs are used to treat serious diseases where a loss of effect can have life-threatening consequences, it is crucial to assess their immunogenic potential.

Some mAbs have been deemed safe and are used to treat autoimmune conditions; however, when applied to other autoimmune diseases, such as multiple sclerosis (MS), they may not be safe despite being approved for marketing. This is due to immunopathogenetic differences between autoimmune diseases, meaning that the immune response triggered by a given mAb may influence the disease differently [7].

It is also important to note the genetic component to immunogenicity, particularly related to an individual's specific Major Histocompatibility Complex (MHC) alleles. Different MHC alleles can bind different peptides derived from the biological drug. In the case of therapeutic mAbs, the mAbs are taken up by antigen-presenting cells, cleaved into peptides, and presented on MHC class II molecules as peptide-MHC (pMHC) complexes. These complexes are transported to the cell surface, where they might bind T-cell receptors (TCRs), thereby initiating downstream immune responses.

Multiple factors must be considered when assessing immunogenicity and the safety of biological drugs, making this a highly complex task. Nonetheless, established methods for

evaluating a drug's immunogenic potential do exist, such as cytokine release assays, including ELISpot IFNγ assays and T-cell proliferation assays [8]. However, the assays can be both costly and time-consuming [8].

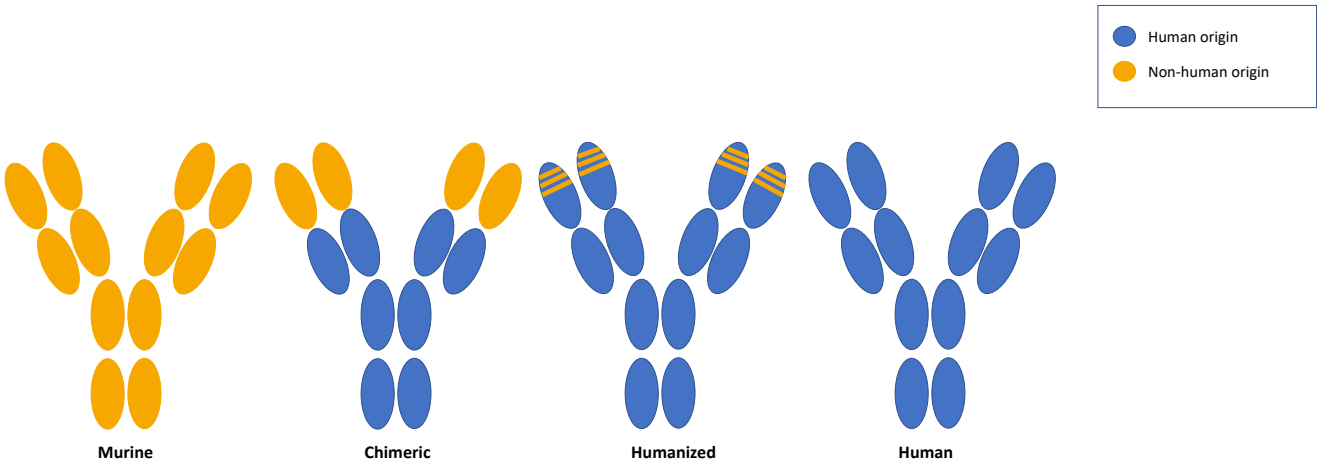


Figure 2: The four main types of antibodies: Murine, chimeric, humanized and human antibodies. Murine antibodies are entirely from non-human origin, whereas chimeric antibodies, the variable region is of non-human origin. In humanized antibodies the complementarity determining regions (CDRs) are of non-human origin and lastly, fully human antibodies.

In silico assessment of toxicity of biological drugs

Several computational approaches have been proposed to assess the immunogenicity of biological drugs. Given the role of humanness in influencing immunogenicity, some tools predict a humanness score for antibodies [9]. The more human-like an antibody is, the less likely it is to be immunogenetic [9]. Other approaches, such as NetMHCIIpan aim to predict the binding affinity to specific MHC alleles [10]. Epimatrix estimates immunogenicity of a protein by predicting T cell epitopes and comparing their frequency to what would typically be expected in a protein of similar length [11]. This comparison yields an immunogenicity score, a score of zero represents the average number of epitopes found in a set of random protein sequences [11]. Scores above zero suggest a higher immunogenic potential, while scores below zero indicate a lower potential [11].

Although several methods exist to facilitate the prediction of the immunogenicity of biological drugs, limited knowledge exists regarding the factors that influence immunogenicity at later stages of the immune response cascade. Moreover, current prediction tools may not account for all factors that influence immunogenicity. Some methods, such as Epimatrix, are commercially available, while others are freely accessible. However, many of these tools focus primarily on early immune recognition events — such as peptide–MHC binding — and do not account for downstream immunological processes, limiting their applicability for comprehensive immunogenicity assessment.

Chapter 2: Methodology background

Data mining and Databases

There are several databases that contain side effect data for biological drugs. As one of these databases distinguish between adverse events (AEs) and adverse drug reactions (ADRs), the neutral term "side effect" is used throughout this thesis. An AE may result from inappropriate use of the drug or may be unrelated to the drug itself [12], whereas an ADR is a harmful and unintentional response that occurs under normal conditions of use, implying a causal relationship between the drug and the side effect [13]. This chapter provides an overview of the data sources and methods used throughout this thesis. More detailed descriptions of how they were applied are provided in the methods section of each study in the results chapter.

The process of analyzing data in various ways to identify patterns and extract meaningful insight is known as data mining [14]. It can be applied throughout the drug discovery pipeline to address specific questions by retrieving and leveraging relevant information from databases. Three different databases were utilized in this thesis — two containing side effect data for therapeutic mAbs, namely PharmaPendium and CLARITY [15], [16], and one, the Immune Epitope Database (IEDB), containing assay data relevant to immunogenicity. In contrast to IEDB, which is publicly accessible, PharmaPendium and CLARITY are commercial databases. As these databases are well curated and were accessible via collaboration partners, this thesis primary focused on the utilization and analysis of the data retrieved from them.

Medical Dictionary Regulatory Activities (MedDRA)

The Medical Dictionary for Regulatory Activities (MedDRA) is a standardized medical terminology developed by the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH) [15]. MedDRA is used internationally, thereby facilitating the exchange of regulatory information across countries [15]. MedDRA uses a hierarchical structure that describes side effects from very specific terms to more general ones, with increasing generality at higher levels of the hierarchy [15]. Figure 3 presents the five hierarchical levels used in MedDRA, listed in order from the most general to the most specific: System Organ Classes (SOCs), High Level Group Terms (HLGTs), High Level Terms (HLTs), Preferred Terms (PTs) and lastly, the most detailed

level, the Lowest Level Terms (LLTs). However, in the databases used in this thesis, side effects were reported using the PTs. Table 1 provides a summary of how the different levels are grouped.

Table 1: The table summarizes how the different MedDRA hierarchical terms are described [15].

MedDRA Hierarchical Terms	**Description of how the terms are grouped**
System Organ Classes	Aetiology, manifestation site or purpose.
High Level Group Terms	Anatomy, pathology, physiology, aetiology or function.
High Level Terms	Anatomy, pathology, physiology, aetiology or function.
Preferred Terms	Symptom, sign, disease diagnosis, therapeutic indication, investigation, surgical or medical procedure, and medical social family history characteristic.
Lowest Level Terms	Observations reported in practice.

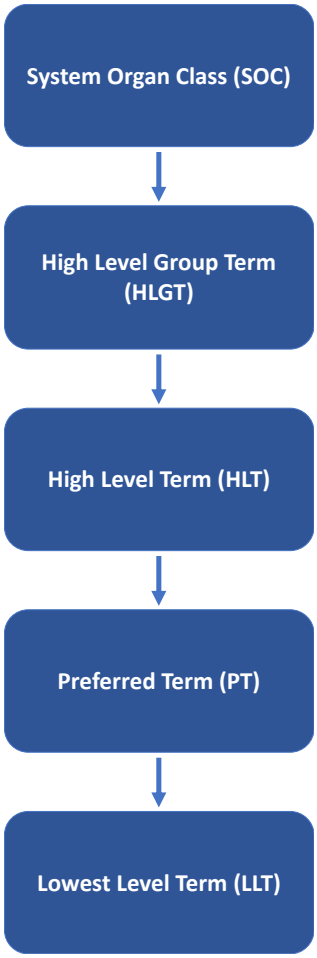


Figure 3: The figure illustrates the hierarchical structure of the MedDRA terminology, beginning with the highest level, the System Organ Class, and ending with the lowest level, Lowest Level Term.

PharmaPendium

The PharmaPendium database contains data for Food & Drug Administration (FDA) and European Medicines Agency (EMA) approved drugs [16]. The data is retrieved from FDA and EMA approval packages, FDA Advisory Committee Meeting documents, FDA Adverse Event Reporting System (FAERS), FDA classic collection (year 1938-1991), Drug Efficacy Study Implementation (DESI) documents, Meyler's 16th Edition, Mosby's Drug Consult, and scientific articles [16]. PharmaPendium provides preclinical, clinical and post-marketing data. In PharmaPendium, if a side effect has been reported once for a certain drug, then the side effect is annotated.

CLARITY

The CLARITY database contains post-marketing data for a wide range of drugs that have already been approved by various regulatory authorities. In CLARITY, the Proportional Reporting Ratio (PRR)— a statistical metric used for signal detection [17], is calculated for every reported side effect and drug. The PRR reflects the proportions of a specific side effect for a given drug in comparison to all other drugs in the database [17]. A PRR greater than 1, suggests that the side effect is reported more frequently for the given drug relative to other drugs. The higher the PRR, the stronger the signal [17], therefore CLARITY applies a threshold of 2 for a side effect to be considered an AE. In addition, CLARITY also annotates whether there is a casual relationship between a specific drug and AE by labelling them as an ADR. Therefore, post-marketing data from CLARITY was used in this thesis due to its high level of curation and the availability of pre-calculated statistical metrics such as the PRR and the annotation of casual relationship of side effects.

Immune Epitope Database (IEDB)

The Immune Epitope Database provides experimental data studied in humans and other species and provides B-cell, T-cell and MHC binding assays. Any researcher can submit data to IEDB. For each peptide it provides a qualitative measurement of the activity in addition to the MHC allele information. When it comes to T-cell assay data, there are different types of T-cell assays with different biological activity.

Machine Learning

Traditional machine learning algorithms were applied for a binary classification task due to the limited amount of data, and several models were explored. Given the imbalanced dataset, various techniques were also employed to address these challenges. The algorithms and techniques used are presented below.

Algorithms

Random Forest

A Random Forest (RF) algorithm is made up of multiple decision trees which in the end gives on single output. A decision tree is comprising of nodes which splits the data based on certain requirements, and in the end helps to arrive at a final decision. However, a decision tree is prone to overfitting. By using multiple decision trees, that capture different patterns in the data, the model can generalize better, resulting in a more accurate predictions. [18]

Support Vector Machine

Support Vector Machines (SVMs) can be used for both linear and non-linear data, depending on the kernel function applied. When a non-linear kernel is used, it maps the data into a high-dimensional space, allowing it to become linearly separable. There are various non-linear kernel functions, and the choice depends on the nature of the data and the specific task. [19]

Gradient Boosting Machine

Another algorithm that consists of a collection of multiple decisions trees is the Gradient Boosting Machine (GBM). However, it differs in that GBM builds model iteratively, and each new tree aims to correct the mistakes by the previous one. This approach can enhance model performance by focusing on the more challenging data points. [20]

Feature Representation

There are various ways to represent molecular structures as input features for machine learning models. In the context of proteins and peptides, the amino acid sequence can be encoded in several ways. One common approach is one-hot encoding, where each amino acid at a given position is represented as a binary vector indicating the presence (1) or absence (0) of that specific residue. While straightforward, this method often results in high-dimensional and sparse representations. An alternative is letter-based encoding (LBE), in which each amino acid is assigned a unique numerical identifier, maintaining the sequential order of the original sequence. This approach yields a more compact representation compared to one-hot encoding. Beyond categorical representations, physicochemical descriptors can be used to capture intrinsic properties of amino acids. For example, Z-scale descriptors quantify various attributes such as polarity, hydrophobicity, and size [21]. These types of descriptors are frequently applied in proteochemometrics, a modelling approach that incorporates information from both the ligand and the protein target, often focusing on binding site characteristics [22]. In addition, recent approaches leverage transfer learning using large pretrained models, which have been trained on extensive datasets of similar molecular structures [23]. These models can extract meaningful features that are then fine-tuned or applied to specific downstream tasks using smaller datasets. Such methods are particularly advantageous when training data is limited, as they help improve generalizability and performance by incorporating prior knowledge.

Feature importance

A feature importance analysis allows for the evaluation of which descriptors are most influential for the model. This information can be used to reduce the number of input descriptors while maintaining, or potentially improving, model performance by reducing noise. Moreover, the provided insights into which types of information the model considers relevant for making predictions can aid in model validation. This is done by assessing whether its reasoning aligns with domain knowledge.

Metrics for Model Evaluation

For binary classification using an imbalanced dataset the following metrics can be used: balanced accuracy (BA), Matthews correlation coefficient (MCC), precision, sensitivity, and specificity (see Equation 1-5). These metrics are calculated based on the number of correctly predicted positives, true positives (TP); correctly predicted negatives, true negatives (TN); incorrectly predicted positives, false positives (FP); and incorrectly predicted negatives, false negatives (FN).

$$BA = \frac{Sensitivity + Specificity}{2} \tag{1}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{4}$$

$$Specificity = \frac{TN}{TN + FP} \tag{5}$$

Active Learning

Active learning (AL) is an approach that begins with a small set of labeled data and iteratively select the most informative unlabeled data points to be labeled and added to the training set. These data points are typically chosen based on specific criteria, such as model uncertainty— for instance, if the model is least confident about certain predictions, those data points are prioritized for labelling. By focusing on the most informative samples, AL can improve model performance more efficiently. This approach is valuable in scenarios where labeled data is scarce or expensive to obtain, as it reduces the amount of data needed to train a model.

Techniques for small and imbalanced datasets in machine learning

In addition to using an AL approach to improve model performance on small and imbalanced datasets, there are several other techniques that can enhance model development. Performing a stratified split during dataset partitioning ensures that class distribution is preserved across training and test sets, which is essential for enabling the model to learn from the minority class while retaining the ability to evaluate its performance as reliably as possible. Hyperparameter optimization is essential for identifying parameter settings that improve model generalizability. This is particularly important when working with a small dataset, which are prone to overfitting, and with imbalanced datasets, where the model may learn more about the majority class while neglecting the minority class. As a result, the model may overpredict the majority class and underperform on the minority class due to lower confidence. In such cases, adjusting the decision threshold for the binary prediction can help to capture the minority class.

Chapter 3: Aim

The main aim of this thesis was to explore toxicity prediction of biological drugs by utilizing data from available databases, analyzing the data, and investigating different ways to describe these macromolecules for machine learning.

The specific research questions were:

1. What do the side effect profiles of biological drugs look like?
2. How predictive are animal models for toxicity assessment of biological drugs?
3. Can a model be built to predict the toxicity of biological drugs, and which descriptors are the most informative?

These questions were addressed in the following studies:

The aim of Study 1 and 2 was to analyze data retrieved form PharmaPendium and CLARITY for therapeutic mAbs, using both animal and human data. Therefore, in Study 1, an interactive dashboard was developed that uses the PTs for any drug and visualize the data for the different MedDRA hierarchies. In addition to the overall side effect analysis, a specific objective was to investigate a side effect that was of interest by our collaboration partners at Sanofi, which was neutropenia, and to assess whether any factors could potentially predict its occurrence. Furthermore, the predictive performance of animal models for neutropenia was evaluated.

The aim of Study 3 was to develop a model for immunogenicity assessment based on sequence data and exploring various descriptors for predictive modelling.

Chapter 4: Results

Study 1: PharmAverse: An Interactive Dashboard Using MedDRA Hierarchy

Abir Omran, Barbara Füzi, Alexander Amberg and Gerhard F. Ecker

Application note submitted, 2025

Abir Omran wrote the python scripts, processed the data, and wrote the application note.

bioRxiv preprint doi: https://doi.org/10.1101/2025.06.17.660110; this version posted June 18, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

PharmAverse: An Interactive Dashboard Using MedDRA Hierarchy

Abir Omran[1], Barbara Füzi[2], Alexander Amberg[2], Gerhard F. Ecker[1*]

[1]University of Vienna, Department of Pharmaceutical Sciences, Josef-Holaubek-Platz 2, Vienna, Austria
[2]Sanofi, Preclinical Safety, Industriepark Höchst, 65926 Frankfurt am Main, Germany

* Correspondence: gerhard.f.ecker@univie.ac.at

Abstract

The Medical Dictionary for Regulatory Activities (MedDRA) is a standardized international medical terminology widely used in databases that collect adverse event data for drugs. MedDRA employs a hierarchical structure, allowing for analysis at different levels of granularity. However, as the volume of adverse event data continues to grow, and given MedDRA's complexity, it becomes increasingly challenging to analyze, and visualize the information effectively. To address this, PharmAverse was developed—an interactive dashboard designed to facilitate the analysis and visualization of adverse event data using MedDRA terms. It offers three main views: Adverse Events, Drug, and Target.

Introduction

Various databases containing adverse event data for drugs, such as the FDA Adverse Event Reporting System (FAERS) and PharmaPendium, use a standardized international medical terminology known as the Medical Dictionary for Regulatory Activities (MedDRA) terminology [1], [2]. MedDRA serves as the ontology for adverse events and follows a hierarchical structure consisting of five levels. The highest level is the "System Organ Classes", followed by "High Level Group Terms", "High Level Terms", "Preferred Terms" and finally, "Lowest Level Terms". The Lowest Level Terms are the expressions used in practice to report observations. However, in databases such as FAERS, Preferred Terms are used for adverse event reporting. By utilizing the different hierarchical levels, one can better understand how the various adverse events are related. It also enables a more global examination of a drug's adverse events by analyzing how the drug affects different system organs. Due to the hierarchical structure and depending on the dataset size, it can be challenging to manually analyze the data efficiently. Therefore, we developed PharmAverse, a dashboard that utilizes MedDRA's hierarchy to visualize adverse event data through three views: Adverse Events, Drug, and Target. This tool is easily integrable and supports the use of preclinical, clinical, and post-marketing data. Furthermore, this tool can be used with adverse event data for any type of drug from any database that utilizes MedDRA terminology. Additionally, the plots can be downloaded for any downstream application. As a use case, we applied the dashboard to therapeutic antibody data to demonstrate its visualization capabilities.

bioRxiv preprint doi: https://doi.org/10.1101/2025.06.17.660110; this version posted June 18, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Features

General Features

The dashboard is implemented in Python using Dash, a framework for building machine learning and data science web applications. The dashboard includes three views: Adverse Events, Drug, and Target. Each view features a dropdown menu that allows the user to select from the available data types—preclinical, clinical, or post-marketing—depending on what has been provided by the user. Additionally, the user can modify the plot size by selecting the desired plot through the dropdown menu on the right side and adjusting the width and height using the sliders under the dropdown menu. If a bar plot is selected, the font size of the x-axis labels can also be adjusted. All plots can be saved as PNG files by clicking the camera icon in the upper right corner of each plot. Figure 1 illustrates the Adverse Events view and highlights the features available across all views (in black) and the features unique to the Adverse Events view (in blue).

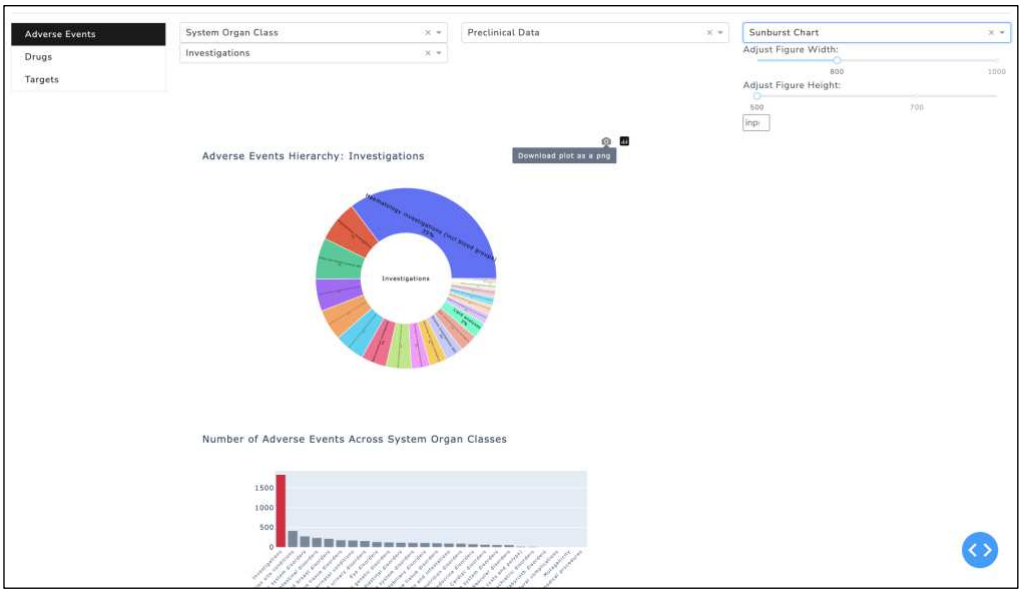


Figure 1: Overview of the Adverse Events view in the dashboard. Features available across all views are marked in black text, while features specific to the Adverse Events view are highlighted in blue. The figure also indicates which plots are interactive. The tabs on the left allow users to navigate between the available views.

Adverse Events View

The Adverse Events view is the initial view displayed upon opening the dashboard — it enables users to analyze the data according to the hierarchical levels defined by MedDRA. This view includes three interactive plots: The first plot is a sunburst chart that displays lower-level adverse event terms linked to the selected System Organ Class from the dropdown menu (see Figure 1). The second plot is a bar chart that visualizes the count of

bioRxiv preprint doi: https://doi.org/10.1101/2025.06.17.660110; this version posted June 18, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Preferred Terms for each System Organ Class, highlighted only the selected class. Users can interact with the sunburst chart by selecting an adverse event term, which will then update both pie chart and the bar chart. If preclinical data is used, a pie chart appears showing the species that reported Preferred Terms associated with the selected adverse event term. In contrast, if clinical or post-marketing data is used, a bar chart will be displayed, which visualizes all drugs reporting the selected adverse event term.

Drug View

The Drug view enables the user to analyze data based on the selected drug. Depending on the available study data, up to four plots can be displayed in this view. The interactive bar chart shows the count of Preferred Terms for each System Organ Class associated with the selected drug. The user can select a System Organ Class from the bar chart, which will update the sunburst chart to display its lower-level terms. Additionally, the sunburst chart is interactive, allowing the user to select an adverse event term at any level to visualize the species in the pie chart which is reporting the adverse events for the selected term when preclinical data is displayed. The pie chart allows the user to select a species, updating the bar chart above to show the count of adverse events (in Preferred Terms) associated with the selected species. When clinical or post-marketing data is used, the bar chart visualizes the counts of Preferred Terms for all drugs reporting the adverse event terms selected from the sunburst chart.

In Supplementary Figure 1, the adverse events for adalimumab are visualized using preclinical data. The System Organ Classes are presented in the bar chart, with the counts for each class displayed. By selecting a System Organ Class, in this case, “Investigation”, the adverse events related to that class are shown in a hierarchical manner in the sunburst chart. Since preclinical data is selected, the pie chart shows the proportion of species reporting the adverse event selected from the sunburst chart, in this case, “White blood cell analyses”. Selecting a species from the pie chart updates the bar chart to display the adverse events associated with the chosen species for all drugs and adalimumab, which is the selected drug is highlighted in red.

Target View

The Target view allows the user to analyze data based on the selected target. This view includes three available plots. By selecting a target from the dropdown menu, the sunburst chart displays the drug(s) associated with the chosen target and indicates whether the drug(s) are linked to other targets. The bar chart visualizes the count of adverse events (in Preferred Terms) for the selected target across all System Organ Classes. Additionally, the user can select a System Organ Class from the bar chart, which updates the second sunburst chart. The second sunburst chart then visualizes the lower-level adverse event terms associated with the chosen System Organ Class.

The Supplementary Figure 2 illustrates the Target view for Tumor Necrosis Factor-alpha (TNF-α). The upper sunburst chart displays all the TNF-α-targeting drugs included in the dataset. The interactive bar plot shows the adverse events counts reported for these drugs across all System Organ Classes. By selecting the System Organ Class “Infections and infestations” from the bar plot, the lower sunburst chart will update to display the MedDRA term hierarchy corresponding to the “Infections and infestations”.

bioRxiv preprint doi: https://doi.org/10.1101/2025.06.17.660110; this version posted June 18, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data Structure and Requirements

A MedDRA license is required to access the MedDRA release package. Begin by downloading the MedDRA release package and place the MedAscii folder in the working directory, this folder contains the MedDRA terminology. The command below generates a dataset that maps MedDRA terms to their respective hierarchical levels. The resulting file will then be used in the data processing step.

```
python meddra_hierarchy_mapping.py
```

To prepare the data for use in the dashboard, ensure that the dataset contains two columns: the first column should represent the drug (e.g., adalimumab), and the second should contain the preferred terms according to MedDRA (e.g., Spleen disorder). For preclinical data, the third column should include species information (e.g., Cynomolgus monkey). The following command processes the data, saves it to the designated folder, and prepares it for integration into the dashboard.

```
python data_processing.py --file_path <path_to_the_file.txt> --
study_type <preclinical | clinical | postmarketing>
```

Additionally, for the Targets view, an extra dataset is required, containing drugs in the first column and their corresponding targets in the second column. This target dataset should be placed in the datasets folder under the name “target_data.csv”.

Availability and Implementation

The project folder containing the source code is available here:
https://github.com/PharminfoVienna/PharmAverse

To launch the dashboard, navigate to the project folder and run the following command in the terminal:

```
python app.py
```

Then, open your web browser and go to the following address:

http://127.0.0.1:8050/

This will launch the dashboard locally on your machine.

bioRxiv preprint doi: https://doi.org/10.1101/2025.06.17.660110; this version posted June 18, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Acknowledgements

In this work the MedDRA® terminology used which is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).

Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) W1232. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. The study was also funded by Sanofi.

bioRxiv preprint doi: https://doi.org/10.1101/2025.06.17.660110; this version posted June 18, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

References

[1] C. for D. E. and Research, 'FDA's Adverse Event Reporting System (FAERS)', FDA. Accessed: Jun. 18, 2025. [Online]. Available: https://www.fda.gov/drugs/surveillance/fdas-adverse-event-reporting-system-faers

[2] 'PharmaPendium | Drug safety, efficacy & DMPK data | Elsevier', www.elsevier.com. Accessed: Jun. 12, 2023. [Online]. Available: https://www.elsevier.com/products/pharmapendium

Study 2: Exploring the Side Effects of Therapeutic Monoclonal Antibodies with Focus on Neutropenia

Abir Omran, Barbara Füzi, Alexander Amberg and Gerhard F. Ecker

Manuscript submitted, 2025

Abir Omran processed the data, performed the analysis, used statistical methods, and wrote most of the manuscript.

Exploring the Side Effects of Therapeutic Monoclonal Antibodies with Focus on Neutropenia

Abir Omran[1], Barbara Füzi[2], Alexander Amberg[2], Gerhard F. Ecker[1*]

[1]University of Vienna, Department of Pharmaceutical Sciences, Josef-Holaubek-Platz 2, Vienna, Austria
[2]Sanofi, Preclinical Safety, Industriepark Höchst, 65926 Frankfurt am Main, Germany

* Correspondence:
Corresponding Author
gerhard.f.ecker@univie.ac.at

Keywords: Monoclonal Antibodies$_1$, Adverse Events$_2$, Adverse Drug Reactions$_3$, Neutropenia$_4$, Drug Safety$_5$

Abstract

Therapeutic monoclonal antibodies (mAbs) are used to treat a wide range of diseases, including cancer, autoimmune disorders, and infections. Known for their high specificity to their targets, they reduce the risk of off-target side effects compared to small molecule drugs. However, mAbs can still cause side effects that are not fully understood. Side effects are classified as adverse events (AEs) or adverse drug reactions (ADRs). AEs may not be drug-related and can arise from improper use, while ADRs are harmful reactions under normal use, implying a causal link to the drug. To better comprehend the diversity of these side effects, and the difference between AEs and ADRs, it is crucial to retrieve and analyze side effect data for mAbs. Therefore, this study aimed to gain a clearer insight into the side effects landscape of mAbs by mining existing data and exploring the difference between AEs and ADRs. Neutropenia, a known side effect of cancer treatments, was used as a case study to investigate whether the treatment type influences the classification of neutropenia as either an AE or an ADR, as well as what information could be retrieved from preclinical data regarding neutropenia. Results show that the therapeutic area combined with target information plays a key role in characterizing the overall ADR profile of therapeutic mAbs. Moreover, our findings indicate that irrespective of the specific disease, the therapeutic area alone provides a more substantial basis for predicting neutropenia classified as an AE than preclinical data.

Introduction

Biological drugs such as therapeutic monoclonal antibodies (mAbs) have gained prominence as therapeutics due to their high target specificity. Consequently, in comparison to small molecules, they cause less direct structure-related and off-target related adverse events (AE). Nevertheless, also for mAbs the possibility for off-target activity exists [1]. For example, abciximab, which targets glycoprotein IIb/IIIa to reduce blood coagulability, is associated with off-target related side effects resulting in hypersensitivity—most likely due to its chimeric origin [2]. Other side effects, like drug-induced haematological disorders such as neutropenia, are attributed to a specific treatment rather than a distinct target. Neutropenia is known to be caused by selected therapeutic mAbs and chemotherapy agents that are used in cancer treatment. Neutropenia is a disorder characterised by neutrophil counts below 1500 neutrophils per microliter of blood, which can be caused by several mechanisms. For instance, the drug could either decrease the production of neutrophils in the bone marrow or increase the rate of cell death of neutrophils. Many anticancer drugs cause bone marrow suppression, which consequently results in neutropenia. However, there are drugs linked to neutropenia despite not being used in cancer treatment [3]. In many of these cases the mechanism of toxicity is unknown.

In this work, we intended to acquire a clearer insight into the side effects of antibodies by mining already existing data. This was conducted by analysing the landscape of all side effects and by completing case studies about neutropenia. Several databases are reporting side effects of drugs reported in preclinical, clinical, and post-marketing studies. These databases collect, store, and structure data to facilitate retrieval and further analysis of side effects. However, the structure and data presentation are very different across these databases. Some databases do not only extract the side effects reports of drugs, but also conduct various statistical analyses which may differ between data sources. Therefore, the databases are insightful in their respective manners and can contribute with varied information in elucidating antibody—side effect connections. PharmaPendium is a database that contains information on preclinical, clinical, and post-marketing reports for drugs approved by the Food and Drug Administration (FDA) and the European Medicines Agency (EMA) [4]. Meanwhile, additional databases such as CLARITY have a broader reporting system and report also drugs awaiting approval or which have been approved by additional agencies [5], [6]. Both databases were included in our study.

Specifically, the focus of this work was to investigate the distinctions of adverse events (AEs) and adverse drug reactions (ADRs), of the side effect landscape of therapeutic mAbs. AEs are not necessarily related to the administrated drug [7] and can occur due to inappropriate use of the intended drug. ADRs, on the other hand, are harmful or unintentional responses to the intended drug [8] that can occur during normal conditions of use. They are often referred to as treatment-related or compound-related findings or reactions. Consequently, with ADRs, there is a causal relationship between the side effect and the administered drug. Considering that drugs used for cancer treatment are known to cause neutropenia, this side effect was used as a use case to examine if the treatment type influences the classification of neutropenia as either ADR or AE. In addition, an animal-to-human translation was conducted to investigate how well preclinical data could capture neutropenia and the differences between neutropenia when classified as an ADR or an AE. Finally, given that ADRs are side effects with a causal relationship, an investigation was conducted to compare the target classes found in the ADRs with those identified in the preclinical data.

Method

Dataset

The first step of the side effect landscape analysis of therapeutic mAbs was to extract data from the CLARITY database. For connecting mAbs to side effects, the proportional reporting ratio (PRR) was utilized, which was determined from post-marketing data for each side effect reported. PRR is a statistical value for signal detection [9]. It conveys the proportions of a specific side effect for the drug of interest compared to all other drugs in the database [9]. A PRR greater than one suggests that the side effect is reported more frequently for the drug of interest in comparison to all the other drugs, and the higher the PRR value, the stronger the signal [9].

For this reason, CLARITY suggests using a PRR threshold of 2 for receiving drug-related associations. Hence, for a side effect to be classified as an AE, the PRR had to be above 2. In addition to the PRR, the causal relationship between drug and side effect is already annotated in ClARITY, meaning if it's an ADR, the annotation provided by the database was directly utilized in our study. The relevant Medical Dictionary for Regulatory Activities (MedDRA) terminologies were used to extract data for the neutropenia analysis.

Medical Dictionary for Regulatory Activities terminology for side effects

All utilized databases use the MedDRA terminologies as the ontology for side effects, thus allowing for combining data from all three data sources. MedDRA uses a hierarchical structure with five levels, with the highest level being the System Organ Classes (SOCs). The subsequent tier is the High Level Group Terms (HLGTs) which is followed by the High Level Terms (HLTs), Preferred Terms (PTs) and lastly the Lowest Level Terms (LLTs). The lowest level terms are the terms that are used in practice to report an observation. However, in the databases used in this work, the PTs are used as the lowest and most detailed level for the side effects. The MedDRA hierarchy for neutropenias is presented in Supplementary Figure 1, with some of its PTs.

Neutropenia

Neutropenia was used as a use case to explore what valuable information can be observed and whether the treatment type, i.e. if the mAb is used for cancer treatment, can influence the classification of neutropenia as either AE or ADR. Here, we used the high-level term neutropenias, which describes various types of neutropenia as the investigated AE and ADR. The post-marketing data used in this study was obtained from the CLARITY database. Finally, the data acquired from this study were employed in the animal-to-human translation study.

Animal-to-human translation

Preclinical data was extracted from PharmaPendium, and we labelled the mAbs neutropenic if there was at least one reporting of the drug connected to the side effect in PharmaPendium. Otherwise, it was labelled as non-neutropenic. In CLARITY, the mAbs were labelled as neutropenic or non-neutropenic based on the PRR threshold mentioned previously. In total, 89 mAbs were shared between the two databases and included in the neutropenia use case analysis and animal-to-human translation.

Likelihood Ratio

The likelihood ratio (LR) was calculated to assess whether a drug's therapeutic area or evidence of neutropenia in preclinical studies serves as an indicator of neutropenia development in humans. LR is a statistical measure used to assess the strength of evidence for a hypothesis based on sensitivity and specificity. It determines the likelihood of a condition occurring in individuals given a particular test result is positive or negative. There is a LR for positive tests, the positive likelihood ratio (LR+), and one for negative tests, the negative likelihood ratio (LR-). The LR+ and LR- are calculated using the formulas in Equation (1) and (2).

In this study, we tested the following hypotheses:

1. If the drug is classified as an antineoplastic agent according to its ATC level 2 code, would it then cause neutropenia as an AE or ADR?
2. If the drug caused neutropenia in animals, would it then cause neutropenia in humans?

A LR+ > 10 suggests that a positive test result indicates a high probability of observing neutropenia in humans. Conversely, a LR- < 0.1 implies a high decrease in the likelihood of observing neutropenia in humans if the test result is negative [10]. Supplementary Table 1 summarizes the interpretations of LR values as described by Jaeschke et al., 1994 [10]. In the analysis, both LR+ and LR- were calculated to evaluate the test's performance in confirming or ruling out neutropenia.

$$LR+ = \frac{sensitivity}{(1 - specificity)} \quad (1)$$

$$LR- = \frac{(1 - sensitivity)}{specificity} \quad (2)$$

Toxicity Profiles

To enable the analysis of the side effect landscape of mAbs, binary fingerprints were created. The fingerprints are based on the AEs or ADRs for the PTs derived from CLARITY and were generated for each mAb, as illustrated in Supplementary Figure 2. The binary fingerprints are referred to as toxicity profiles, and the value 1 corresponds to a side effect being present, and 0 corresponds to the side effect being absent. The intent was to focus on AEs and ADRs that were reported for at least one other mAb. Therefore, the length of the fingerprints corresponds to the total number of AEs and ADRs extracted from the database, which are 1 827 and 1 828, respectively. The side effect that was present in AE but not in ADR fingerprints was nasal abscess.

Anatomical Therapeutic Chemical code

To quantify and visualize the treatment-related side effects for mAbs in their side effect landscape, we had to characterize them based on their therapeutic area. This was done by using the Anatomical Therapeutic Chemical (ATC) classification system. It divides drugs into different groups according to the organ or system they act on, as well as their therapeutic, pharmacological, and chemical properties. Supplementary Table 2 presents the five ATC levels and a description for each level, while Supplementary Figure 3 provides the ATC classification for the therapeutic mAb rituximab.

We decided to focus on the ATC level 2 code to describe the therapeutic area of the mAbs, which was extracted from the ChEMBL database. The ATC level 3 code was excluded as it was very similar to level 2, therefore being redundant. The ATC level 1 code was too broad, and the ATC level 4 code was too target-specific, at least for some mAbs. If a drug is registered to have more than one ATC level 2 code and at least one is the code for antineoplastic agents— then only the antineoplastic agents label was considered for the mAb.

Clustering

Clustering was performed using the UMAP package in Python. The UMAP algorithm with the Jaccard distance was implemented. The parameters ‘n_neighboor’ and ‘min_dist’ were adjusted until a clear distinction of the groups was obtained. The toxicity profiles were used as input to examine the side effects pattern of the mAbs for both AEs and ADRs and coloured based on their ATC level 2 code.

Results

Analysing All Adverse Drug Reactions and Adverse Events

All side effects connected to mAbs were used to investigate their side effect landscape and to evaluate the potential valuable insight that can be obtained by encoding all the AEs and ADRs into toxicity profiles. The results from clustering the toxicity profiles are visualised in Figure 1 and coloured according to their ATC level 2 code.

The result obtained from clustering the toxicity profiles generated from the ADRs shows that the mAbs were mostly grouped according to their therapeutical information described by the ATC level 2 code. However, several outliers exist, such as the anti-interleukin-23 (IL-23) mAb, Guselkumab, an immunosuppressant located closer to the antineoplastic agents than to the immunosuppressant ones. The other anti-IL-23 mAbs, Risankizumab and Ustekinumab, were located closer to the other immunosuppressants. The anti-tumor necrosis factor alpha (anti-TNF-α) mAbs, Adalimumab, Golimumab, and Infliximab were, by contrast, located relatively close to each other and close to the other immunosuppressants.

In contrast to the result obtained from the ADRs, the AEs, presented in Figure 1b, showed that the clustering based on the therapeutical information was less distinct. Nonetheless, focusing on the anti-IL-23 mAbs, Guselkumab was no longer near the antineoplastic agents. Similarly, as for ADRs, the anti-TNF-α mAbs were still near the immunosuppressant mAbs.

Figure 1: An illustration of the toxicity profiles of the adverse drug reactions (ADRs) and adverse events (AEs) using the Preferred Terms (PTs). A) shows ADRs, while B) shows AEs. The terms are colored based on the ATC level 2 code of the corresponding monoclonal antibodies (mAbs).

Neutropenia

Anticancer drug and neutropenia

Given that anticancer drugs are commonly associated with neutropenia, they were used as a case study to examine whether this is reflected in post-marketing data. Additionally, the analysis compared how neutropenia is reported when it's classified as an AE versus an ADR. The HLT neutropenias was used to extract all mAbs listing it as an AE and ADR. In total 13 mAbs reported neutropenias as an ADR; out of these, eight are antineoplastic agents and five are immunosuppressants. However, regarding the AEs, 33 mAbs reported neutropenias as an AE and out of these, 25 were antineoplastic agents, seven were immunosuppressants and one was an immune sera and immunoglobulins.

The likelihood ratio was calculated to investigate if a drug used for cancer treatment, which in this case means an antineoplastic agent drug, would alter the probability that the patient develops neutropenia due to the drug's therapeutical area. The result presented in Table 1 shows that, when neutropenia was classified as an ADR, the LR+ was 2.60 and the LR- was 0.50. This indicates that there is a small probability to develop neutropenia as an ADR after the use of antineoplastic agent drugs. Alternatively, as exemplified by the LR- value, not taking the drug leads to a small decrease in the likelihood of developing neutropenia. This indicates that the therapeutic area as described by the ATC code is not a strong predictor for neutropenia classified as an ADR. On the other hand, neutropenia classified as an AE resulted in a LR+ above 10, which points towards a high probability that a patient treated with an antineoplastic agent drug could develop neutropenia. Consequently, the therapeutic area of the drug is a strong predictor for neutropenia. The LR- value, however, was above 0.2 meaning that the absence of the drug would lead to a small decrease in the likelihood of developing neutropenia.

Table 1: In this table, the positive and negative likelihood ratios are presented for when neutropenia was classified as an adverse drug reaction (ADR) and when it was classified as an adverse event (AE).

Side Effects	Positive likelihood ratio	Negative likelihood ratio
Adverse Events	42.42	0.25
Adverse Drug Reactions	2.60	0.50

Animal-to-human translation

The result from the animal-to-human translation analysis showed the lack of concordance between preclinical and post-marketing data for neutropenia when classified as AE or ADR. Table 2 displays the calculated LR for both AEs and ADRs. This was performed to investigate whether the likelihood of neutropenia classified as AE or ADR would also be observed in post-marketing studies when already observed in preclinical studies. Although the LR+ for the ADRs was slightly higher than for AEs, they were still below 2. This indicates a low probability of observing neutropenia as an AE and ADR in post-marketing studies if it has been observed in preclinical studies. The LR- was slightly below 1 for both AE and ADR, which indicates that the absence of neutropenia in preclinical studies would lead to a minimal decrease in the likelihood of neutropenia in post-marketing studies.

Table 2: A summary of the likelihood ratio (LR) result for neutropenia when classified as an adverse drug reaction (ADR) or adverse event (AE).

Side Effects	Positive likelihood ratio	Negative likelihood ratio
Adverse Events	1.70	0.97
Adverse Drug Reactions	1.95	0.96

The neutropenic mAbs in the preclinical data were further investigated to compare their pharmacological targets with those of mAbs reporting neutropenia classified as an ADR in the post-marketing data. Four mAbs linked to four distinct targets were obtained from the preclinical data and compared to the eight targets identified for the ADRs (see Supplementary Table 3). As presented in Table 3, only two mAbs have neutropenia as an AE, whereas only one also has it as an ADR. These were Avelumab and Rituximab. Both are antineoplastic agent drugs and are not unique to their target. However, they are the only mAbs from their target group that showed neutropenia in preclinical studies. The two other mAbs that were identified as neutropenic in the preclinical data but not in the post-marketing data are immunosuppressants. Some neutropenic mAbs in the post-marketing data are also immunosuppressants. However, they target different molecules, namely CD52 and interleukin-6 (IL-6) and its receptors (IL-6R) (see Supplementary Table 3).

Table 3: A summary of the target classes identified in the preclinical data, including the monoclonal antibodies (mAbs) associated with each class. The table also indicates whether the target classes were found in the post-marketing data and their respective side effect types. The targets include Programmed death-ligand 1 (PD-L1), Interleukin-4 receptor subunit alpha (IL-4Rα), Integrin α4β1 (VLA-4) and B-lymphocyte antigen CD20.

Target	Monoclonal Antibodies	Type of Side Effect
PD-L1	Avelumab	AE
IL-4Rα	Dupilumab	-
VLA-4	Natalizumab	-
CD20	Rituximab	AE/ADR

Discussion

In this study, the side effect landscape of therapeutic mAbs was investigated, focusing on side effects classified as AEs and ADRs. The analysis aimed to examine the distinctions between AEs and ADRs for therapeutic mAbs, using post-marketing data. The results indicate that, at the PT level, the ADRs are linked to the therapeutic area as described by the ATC level 2 code. Havin a closer look at the therapeutic areas showed that the mAbs were grouped based on their pharmacological targets, as seen for with the anti-TNF-α mAbs (Figure 1). Similarly, also for the AEs the mAbs were grouped based on their pharmacological target. However, for AEs the link to the therapeutic area was less pronounced. This might be due to the use of the ATC level 2 code, which lacks specificity regarding the exact indication (i.e. which kind of cancer the drug is used for). This indicates, that for assignment of ADRs a higher-level description of the use of the drug is sufficient, while for AEs the situation is more complex. Accordingly, the analysis of the toxicity profiles shows that the therapeutic area with the target type of the mAb can be pivotal to understanding and determining its ADRs. However, there are some outliers, as demonstrated by the anti-IL-23 mAbs. They are all located closer together in terms of the side effect landscape of the AEs, compared to how they are positioned in the side effect landscape of the ADRs.

These findings are generally expected from mAbs due to their high specificity. Hence, a use case in which a side effect is linked to a specific treatment type was examined to determine if the findings from the previous analysis align with those of the use case. This led to the study of neutropenia and the likelihood of developing neutropenia as an AE or ADR when the drug in question is used for cancer treatment. This was done by calculating the LR for neutropenia classified as AE and ADR based on the ATC level 2 code, determining whether the drug was categorized as an antineoplastic agent. For this analysis, the HLT neutropenia was used, which is from a higher-tier level compared to the PT level used in the side effect landscape analysis. The result showed that if the drug is an antineoplastic agent, there is a high probability of developing neutropenia classified as an AE. Nonetheless, the LR- indicated that the absence of the antineoplastic agent, in turn, leads to a moderate decrease in the likelihood of developing neutropenia as an AE. This result suggests that the use of antineoplastic agents is a strong indicator of the development of neutropenia. However, when the drug is not classified as an antineoplastic agent, one cannot entirely rule out the condition. In cancer treatment, it is common to use multiple drugs simultaneously and at higher doses. Consequently, it is not unexpected that these medications could cause neutropenia as an AE. Moreover, AEs can also result from side effects unrelated to the administered drug or from inappropriate use of the drug. This might also explain why the side effect profile of AEs cannot be accounted for by therapeutic information to the same extent as ADRs.

While our studies suggest that therapeutic information plays a key role in determining a drug's ADRs, the neutropenia use case did not align with this trend. The use of antineoplastic agents showed only a slight increase in the probability of neutropenia being classified as an ADR. Simultaneously, the LR- value of 0.5 indicates only a modest decrease in the likelihood of developing neutropenia in the absence of these agents. Thus, in this case, therapeutic information serves as a weak predictor for neutropenia as an ADR. Since a causal relationship between the drug and the specific side effect is required to classify it as an ADR, the pharmacological targets of the mAbs causing a neutropenic ADR were further examined. Six mAbs that are classified as immunosuppressants reported neutropenia as an ADR. Four mAbs target the IL-6 ligand or the IL-6 receptor, one targets CD52, and the last one targets CD20. In this context it is worth mentioning that both the IL-6 receptor and CD52 are

expressed on neutrophils. The four monoclonal antibodies (mAbs) targeting IL-6 or its receptor (IL-6R) block the interaction between the IL-6/IL-6R complex and glycoprotein 130 (gp130)-associated Janus kinases (JAKs) [3]. Additionally, there are also small-molecule JAK inhibitors available, which act downstream of gp130 [11]. Similar to anti-IL-6R mAbs, certain JAK inhibitors— such as Tofacitinib—function as immunosuppressants and have been associated with neutropenia [12]. Thus, our findings for mAbs are consistent with previous research, supporting the notion that inhibition of IL-6 signalling or its downstream pathways can contribute to the development of neutropenia [3]. While the use of ADR data at the HLT level for neutropenia appears to identify mechanistically related drugs, it's important to note that this does not imply that the approach can be used to distinguish between on-target and off-target effects.

Finally, an animal-to-human translation study for the high-level term neutropenias was conducted to evaluate the consistency between preclinical and human data. LRs were calculated to determine whether neutropenia— classified as either an AE or ADR—in humans could be predicted based on its occurrence in animals. The results revealed only a slight increase in the likelihood of observing neutropenia in humans when it had been previously observed in animals, and a similarly modest decrease when it had not. Moreover, the comparison of target classes derived from preclinical and post-marketing ADR data showed little overlap. Only one mAb was common to both datasets, while other mAbs targeting the same proteins were not identified in preclinical studies. Notably, anti-IL-6/IL-6R mAbs, which are well-documented to induce neutropenia, were also absent from the preclinical findings. This discrepancy may be attributed to the fact that IL-6 signalling does not play as pivotal a role in immune response in animals as it does in humans [13], [14]. Consequently, blocking IL-6 pathways leads to different side effect profiles across species [14]. In the case of neutropenia, these results suggest that therapeutic area information alone provides a stronger predictive signal than preclinical data.

Conclusion

This study analyzed the side effect landscape of therapeutic mAbs using toxicity profiles derived from all reported AEs and ADRs. Results indicate that the therapeutic area, classified via the ATC level 2 code, combined with target information plays a key role in characterizing the overall ADR profile of therapeutic mAbs. Additionally, our findings show that, regardless of the specific disease, the therapeutic area alone offers a higher probability of predicting neutropenia classified as an AE compared to relying on preclinical data. This analysis contributes to a deeper understanding of curated side effect data for therapeutic mAbs, particularly in distinguishing between AEs and ADRs, with a focus on neutropenia.

Author Contributions

A.O. conducted the majority of the research and wrote the manuscript with support from B.F.. A.A. and G.E. contributed to the design and conceptualization of the study
All authors contributed to the discussions, provided feedback, and reviewed the final manuscript.

Acknowledgements

In this work the MedDRA® terminology used which is the international medical terminology developed under the auspices of the International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use (ICH).

Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) W1232. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. The study was also funded by Sanofi.

References

[1] D. Bumbaca *et al.*, 'Highly specific off-target binding identified and eliminated during the humanization of an antibody against FGF receptor 4', *mAbs*, vol. 3, no. 4, pp. 376–386, Jul. 2011, doi: 10.4161/mabs.3.4.15786.

[2] A. L. Catapano and N. Papadopoulos, 'The safety of therapeutic monoclonal antibodies: Implications for cardiovascular disease and targeting the PCSK9 pathway', *Atherosclerosis*, vol. 228, no. 1, pp. 18–28, May 2013, doi: 10.1016/j.atherosclerosis.2013.01.044.

[3] H. L. Wright, A. L. Cross, S. W. Edwards, and R. J. Moots, 'Effects of IL-6 and IL-6 blockade on neutrophil function in vitro and in vivo', *Rheumatology*, vol. 53, no. 7, pp. 1321–1331, Jul. 2014, doi: 10.1093/rheumatology/keu035.

[4] Elsevier, 'PharmaPendium'. [Online]. Available: https://www.elsevier.com/products/pharmapendium

[5] 'ClarityVista'. Accessed: Jun. 18, 2023. [Online]. Available: https://clarity-vista.com/

[6] 'OFF X'. Accessed: Jun. 19, 2023. [Online]. Available: https://www.targetsafety.info/target-class-scores-information

[7] 'Adverse event | European Medicines Agency'. Accessed: Jul. 16, 2024. [Online]. Available: https://www.ema.europa.eu/en/glossary/adverse-event

[8] 'Adverse drug reaction | European Medicines Agency'. Accessed: Jul. 29, 2024. [Online]. Available: https://www.ema.europa.eu/en/glossary/adverse-drug-reaction

[9] S. J. W. Evans, P. C. Waller, and S. Davis, 'Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports', *Pharmacoepidemiology and Drug*, vol. 10, no. 6, pp. 483–486, Oct. 2001, doi: 10.1002/pds.677.

[10] R. Jaeschke *et al.*, 'Users' Guides to the Medical Literature: III. How to Use an Article About a Diagnostic Test B. What Are the Results and Will They Help Me in Caring for My Patients?', *JAMA*, vol. 271, no. 9, pp. 703–707, Mar. 1994, doi: 10.1001/jama.1994.03510330081039.

[11] S. Rose-John, B. J. Jenkins, C. Garbers, J. M. Moll, and J. Scheller, 'Targeting IL-6 trans-signalling: past, present and future prospects', *Nat Rev Immunol*, vol. 23, no. 10, pp. 666–681, Oct. 2023, doi: 10.1038/s41577-023-00856-y.

[12] T. S. Mitchell, R. J. Moots, and H. L. Wright, 'Janus kinase inhibitors prevent migration of rheumatoid arthritis neutrophils towards interleukin-8, but do not inhibit priming of the respiratory burst or reactive oxygen species production', *Clinical and Experimental Immunology*, vol. 189, no. 2, pp. 250–258, Jul. 2017, doi: 10.1111/cei.12970.

[13] J. Hoge *et al.*, 'IL-6 Controls the Innate Immune Response against Listeria monocytogenes via Classical IL-6 Signaling', *The Journal of Immunology*, vol. 190, no. 2, pp. 703–711, Jan. 2013, doi: 10.4049/jimmunol.1201044.

[14] R. D. Metcalfe, T. L. Putoczki, and M. D. W. Griffin, 'Structural Understanding of Interleukin 6 Family Cytokine Signaling and Targeted Therapies: Focus on Interleukin 11', *Front. Immunol.*, vol. 11, Jul. 2020, doi: 10.3389/fimmu.2020.01424.

Study 3: Exploring Diverse Approaches for Predicting Interferon-Gamma Release: Utilizing MHC Class II and Peptide Sequences

Abir Omran, Alexander Amberg and Gerhard F. Ecker

Briefings in Bioinformatics, 2025, 26(2).

Abir Omran retrieved and processed the data, built the machine learning model, performed the analysis, and wrote the manuscript.

Briefings in Bioinformatics, 2025, **26(2)**, bbaf101
https://doi.org/10.1093/bib/bbaf101
Problem Solving Protocol

Exploring diverse approaches for predicting interferon-gamma release: utilizing MHC class II and peptide sequences

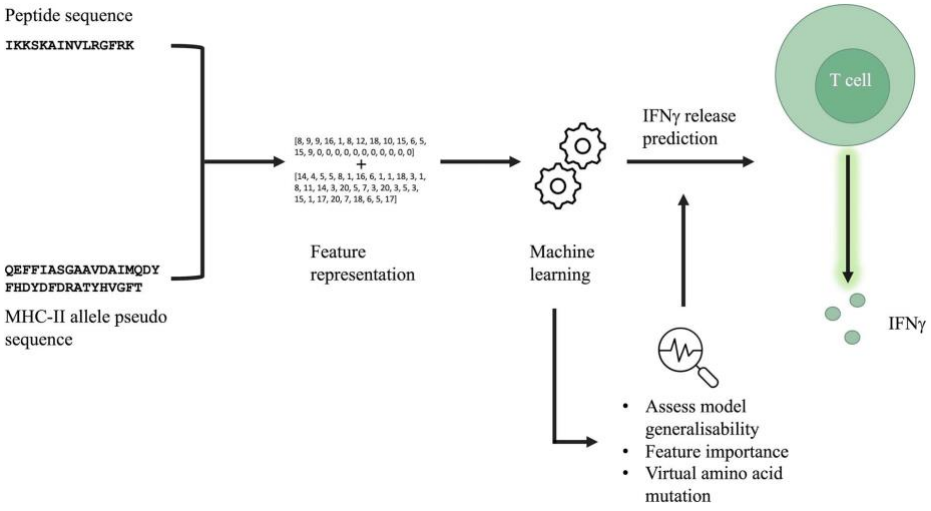
Abir Omran[1], Alexander Amberg[2], Gerhard F. Ecker[1,*]

[1]Department of Pharmaceutical Sciences, University of Vienna, Josef-Holaubek-Platz 2, 1090 Vienna, Austria
[2]Sanofi, Preclinical Safety, Industriepark Höchst, 65926 Frankfurt am Main, Germany
[*]Corresponding author. Department of Pharmaceutical Sciences, University of Vienna, Josef-Holaubek-Platz 2, 1090 Vienna, Austria.
E-mail: gerhard.f.ecker@univie.ac.at

Abstract

Therapeutic proteins are in high demand due to their significant potential, driving continuous market growth. However, a critical concern for therapeutic proteins is their ability to trigger an immune response, while some treatments rely on this response for their therapeutic effect. Therefore, to assess the efficacy and safety of the drug, it is pivotal to determine its immunogenicity potential. Various experimental methods, such as cytokine release or T-cell proliferation assays, are used for this purpose. However, these assays can be costly, time-consuming, and often limited in their ability to screen large peptide sets across diverse major histocompatibility complex (MHC) alleles. Hence, this study aimed to develop a computational classification model for predicting the release of interferon-gamma based on the peptide sequence and the MHC class II (MHC-II) allele pseudo-sequence, which represents the binding environment of the MHC-II molecule. The dataset used in this study was obtained from the Immune Epitope Database and labeled as active or inactive. Among the approaches explored, the random forest algorithm combined with letter-based encoding resulted in the overall best-performing model. Consequently, this model's generalizability to other T-cell activities was further evaluated using a T-cell proliferation dataset. Furthermore, feature importance analysis and virtual single-point mutations were conducted to gain insights into the model's decision-making and to improve the interpretability of the model.

Graphical Abstract



Keywords: machine learning; T-cell response; immunology; predictive modeling

Received: October 4, 2024. **Revised:** February 6, 2025. **Accepted:** February 21, 2025
© The Author(s) 2025. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

Introduction

In 1880, the first antibody-based treatment, a serum therapy against diphtheria, was discovered by Behring [1, 2]. It was not until 1986 that the first therapeutic monoclonal antibody (mAb), Muromonab-CD3, a murine mAb, was approved for clinical use [1]. Muromonab-CD3 frequently caused severe adverse reactions, including cytokine storms, leading to its withdrawal in 2010 [3]. Today, the market size for therapeutic proteins continues to grow and is estimated to reach 47.4 billion US dollars by 2032 [4]. Therapeutic proteins are widely sought-after drugs, offering treatment for numerous acute and chronic diseases that previously were considered untreatable. Compared to small molecules, they exhibit high selectivity, which generally results in fewer off-target adverse events. Nonetheless, therapeutic proteins can still provoke an immune response, potentially leading to a variety of adverse events.

Vaccines rely on the immune response as part of their efficacy. They introduce an antigen into the body, prompting the immune system to respond to it. This process enables the body to 'remember' the antigen, allowing for a faster and more effective defense upon future exposure. A key component of this immunogenic response is the major histocompatibility complex (MHC) presentation pathway, which plays a crucial role in antigen recognition and immune activation. The process starts with the cleavage of the protein into peptide fragments that can potentially bind to the MHC molecules. Peptides binding to MHC class II (MHC-II) usually have a length of around 12–25 residues [5, 6]. The MHC-II molecule contains four binding grooves and specifically recognizes positions 1, 4, 6, and 9 on the peptide, which are essential for binding [5]. In addition, genetic factors influence binding, as peptides interact with different alleles with varying affinities, leading to differences in the immunogenic response [7]. Even a single-point mutation in an MHC-II molecule can significantly alter its binding affinity [8]. When a peptide binds to the MHC-II molecule, it forms a peptide–MHC-II complex (pMHC-II). Subsequently, this complex translocates to the cell surface, where it presents the peptide to a T-cell, initiating T-cell activation. Specifically, the MHC-II molecule presents the peptide to a CD4+ T-cell, also known as a T-helper cell. The peptide residues exposed to the T-cell receptor and primarily involved in binding are positions 2, 3, 5, 7, and 8 [9]. Upon activation, T-helper cells can differentiate into distinct subsets, including T-helper 1 (T_H1) cells. These cells play a crucial role in cell-mediated immunity by secreting cytokines that activate other immune cells, such as cytotoxic T-cells (CTL). The main cytokine secreted by T_H1 cells is interferon-gamma (IFNγ). However, an excessive pro-inflammatory response can result in tissue damage [10] or chronic inflammation [11]. Therefore, assessing a drug's immunogenicity potential is essential to ensure both its safety and efficacy.

According to the Food and Drug Administration guidance on immunotoxicity, the most suitable assay for assessing immunogenicity risk should be selected, such as the cytokine release assay [12]. There are several methods to evaluate if a peptide has the potential to induce an immunogenic response. The ELISpot assay can be used to measure IFNγ release, providing insight into T-cell activation and the immunogenic potential of a therapeutic protein [13]. However, a cytokine release assay requires allele matching, which is crucial for accurate measurement outcomes [14]. MHC-multimer-based flow cytometry is another technique used to assess T-cell response, but due to HLA polymorphism, it is not suitable for population-wide analysis and is only effective for individuals with a specific set of MHC alleles [15]. Other methods, such as T-cell proliferation assays, require prolonged incubation times [14], making them both time-consuming and costly. Consequently, it is essential to develop methodologies capable of rapidly and efficiently screening a large set of peptides across multiple MHC alleles. In this context, *in silico* methods serve as valuable tools, enabling rapid and comprehensive analysis, which is crucial for predicting and assessing a peptide's immunogenic potential for a given MHC allele of interest.

Several *in silico* methods exist for predicting and evaluating a peptide's immunogenic potential. POPISK is a model that predicts whether a peptide associated with the HLA-A02:01 allele has the potential to be immunogenic based on T-cell response data [16]. Other computational tools, such as NetMHCIIpan, focus on predicting peptide binding affinity to a given MHC allele [17]. Additionally, models leveraging Bidirectional Encoder Representations from Transformers (BERT) models have been trained on a high number of protein sequences to address biological tasks, including peptide–MHC binding predictions [18, 19]. BERT models, originally developed for natural language processing, generate rich contextual embeddings from large protein sequence datasets, making them valuable for downstream analyses. Other protein descriptors, such as z-scale descriptors, are commonly used in proteochemometric models to describe the binding site environment of proteins. The descriptors characterize the amino acids' physicochemical properties [20], providing informative input for the model.

Dhanda *et al.* developed an *in silico* method for predicting a specific cytokine release, such as IFNγ [21]. Their approach utilized binary patterns, composition, and motifs as features for the peptide sequence to predict IFNγ release. The model, however, did not account for the MHC-II allele in the prediction. To the best of our knowledge, the model proposed by Dhanda et al. is currently the only model for IFNγ release prediction. Therefore, in this study, we explored various approaches for building IFNγ release models using information from the peptide sequences and the MHC-II allele pseudo-sequences. Given the complexity of the T-cell response, we explored different approaches to enhance the model's performance and generalizability. This involved testing various descriptors, such as letter-based encoding (LBE), ProtBert embedding features, and z-scale descriptors, to characterize both peptide and MHC-II allele sequences. In addition, we adopted an active learning (AL) approach. Beyond evaluating model performance, our study also examined model interpretability through feature importance analysis and virtual single amino acid mutation experiments in the peptide sequences.

Materials and Methods

Data collection and preprocessing

Data were obtained from the Immune-Epitope Database (IEDB) and filtered for a human host, MHC-II, and positive and negative assays. The pMHC-II pairs were labeled based on the majority measurement. For instance, if the pMHC-II had five measurements recorded in the database for IFNγ release, and three out of five were negative, then the pMHC-II was labeled as inactive. Pseudo-sequences representing the binding site environment of MHC-II molecules, each 34 residues in length, were used for the MHC-II alleles [22]. The dataset was further refined based on peptide length, as different studies have suggested varying optimal lengths. For this study, we selected peptides ranging from 12 to 24 residues. Lastly, all duplicates based on the peptide sequence, the MHC-II allele pseudo-sequence, and activity were

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

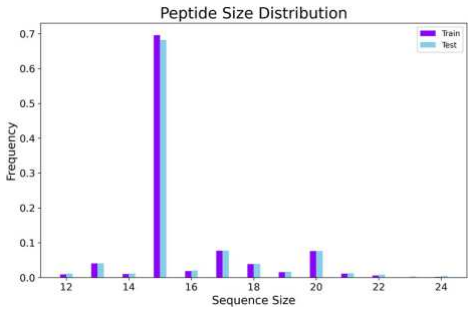


Figure 1. The figure visualizes the peptide sequence length distribution in the training and test set in the first fold. The majority length was 15 residues, accounting for 70% of the dataset.

reduced to a single instance, whereas duplicates based solely on the peptide sequence and the MHC-II allele pseudo-sequence were completely removed from the dataset. This resulted in an imbalanced dataset with 7266 pMHC-II pairs, where the inactive class accounted for 30% of the total data.

Dataset split

A 10-fold cross-validation (CV) was performed, and given the imbalance in classes and peptide sequence lengths, stratified splitting was applied to preserve the original dataset's distribution across all subsets. As shown in Fig. 1, the most common peptide length was 15 residues, which comprised 70% of the dataset.

Model development

Representation

We explored three types of descriptors: LBE, ProtBert embedding features, and z-scale descriptors. While LBE and ProtBert features utilized peptides of various lengths, z-scale descriptors required sequences of uniform length. Therefore, we selected the majority length in the dataset (15 residues), resulting in a 30% reduction in dataset size. The z-scale descriptors were generated using the peptides library in Python. In order to enable the comparison with the LBE method, the same dataset was used, referred to as LBE 15mer.

The ProtBert model was used to generate features for the sequences in Python. Trained on 217 million protein sequences, ProtBert captures contextual information that could be beneficial for our task. [23]. These features have the potential to enhance model performance and generalizability. We started with ProtBert, as it was less complex than some of the other BERT models, allowing us to establish a clear baseline without imposing additional computational demands.

The descriptors used to describe the peptide sequence and the MHC-II pseudo-sequence were concatenated and then used as input to the model. However, before concatenating the LBE descriptors, the peptide sequences were padded with zeros until reaching the length of 25, as depicted in Fig. 2.

Model building

Classical machine learning algorithms, such as random forest (RF), provide greater interpretability than deep learning models. Thus, RF, support vector machine (SVM), and gradient boosting machine (GBM) were selected as the three algorithms used in this study across the three different feature representations. To address class imbalance in the dataset, the probability threshold for classifying active samples was adjusted. Rather than using the default threshold of 0.5, multiple values were tested, and the threshold that achieved a better balance between sensitivity and specificity was chosen for the final model, which was 0.65. Another approach to address the imbalanced dataset was AL [24], applied to the LBE model. For each iteration, 10 samples with the highest uncertainty were added to the training set to refine the model. This approach helps the model to learn the data distribution where it struggles most, meaning that in the case of imbalanced data, the model might gain more insight into the minority class. As a result, AL has the potential to enhance generalization. The best-performing algorithm for the LBE model was subsequently used for the LBE 15mer model and the AL approach.

Hyperparameter optimization

A randomized search was performed instead of a more exhaustive grid search, as it allows for a more efficient exploration of a wider range of hyperparameters. This approach provides better coverage in comparison to a grid search with the same parameter settings. The random search was performed with a 10-fold CV for all models, with the selected parameters detailed in Supplementary Tables 1–2. The same hyperparameters found for LBE were also utilized for the AL approach.

Model performance evaluation metrics

For evaluation of the performance of the binary models, the scikit-learn library was employed in Python to calculate various metrics. The metrics included balanced accuracy, Matthews correlation coefficient (MCC), precision, sensitivity, and specificity.

Model assessment

T-cell proliferation dataset

Since no additional IFNγ release data could be identified, creating an external test set was not feasible. To address this limitation, an alternative dataset was used to evaluate if and to what extent the model could predict the outcome of a T-cell assay type closely associated with IFNγ release. Data points for T-cell assay types, such as T-cell proliferation and IL-2 release, were extracted from IEDB. After filtering out overlapping data from the IFNγ release dataset, a T-cell assay with sufficient data containing both active and inactive samples was obtained. The final T-cell proliferation dataset comprised 711 data points, with 600 active and 111 inactive samples, making it suitable for prediction. All other assay types provided a smaller dataset with very few inactive samples. The T-cell proliferation dataset was applied to the model that demonstrated the best overall performance and was used in place of a test set in the 10-fold CV, enabling a direct comparison with the performance on the IFNγ release test set.

Model interpretability

Feature importance

A feature importance procedure was conducted to identify the top five most important amino acid positions. Subsequently, the amino acid frequency at these positions was examined to identify discrepancies between the two classes, helping to assess whether the model was capturing valuable information regarding T-cell response. The feature importance was only performed on the LBE 15mer model to facilitate the analysis and interpretability. The scikit-learn library was used to compute the feature importance for tree-based models.

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

Peptide sequence	MHC-II pseudo sequence	Allele	Activity
IKKSKAINVLRGFRK	QEFFIASGAAVDAIMQDYFHDYDFDRATYHVGFT	HLA-DRB5*01:01	Positive
TAGVIVMLIPTVMAF	QEFFIASGAAVDAIMESGFDYYDFDRLTYHVGFT	HLA-DRB1*08:02	Positive
RILIGFLVLWIGTNS	QEFFIASGAAVDAIMESSFDYFDFDRATYHVVFT	HLA-DRB1*11:04	Negative

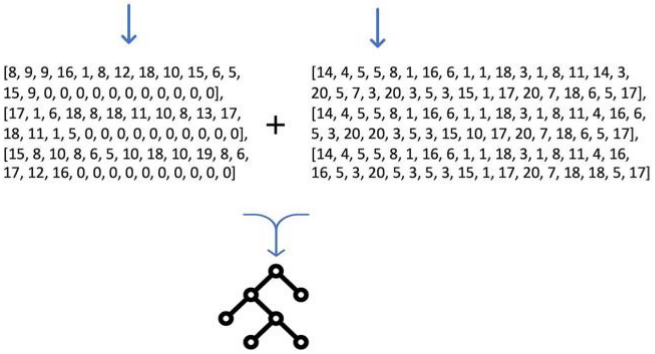


Figure 2. The figure depicts the data for the modeling and how the data was used as input for the LBE models. Unique numerical values representing the amino acids were used to encode the sequences. Subsequently, the LBE was concatenated and used as input for the model.

Virtual peptide single amino acid mutation

A virtual single amino acid mutation was applied to each distinct amino acid at every position individually within the test set. This approach allowed for the assessment of which amino acids in specific positions had a greater impact on the model's predictions. Therefore, the difference in error rate (ΔER) between the true prediction against the mutated prediction was used as a metric in this analysis, calculated as shown in equations (1) and (2). A ΔER value of zero indicates that the mutation had no effect on the prediction, a positive ΔER value signifies an increase in ER, while a negative ΔER reflects a decrease in the ER. For consistency with the feature importance analysis, this method was applied exclusively to the LBE 15mer model.

$$Error\ rate = \frac{Number\ of\ incorrect\ predictions}{All\ predictions} \quad (1)$$

$$\Delta Error\ rate = Error\ rate_i – Error\ rate_j \quad (2)$$

The predictions were performed twice to calculate the ΔER, and overlapping mutations that resulted in a statistically significant ΔER within the top five positions were selected for further analysis. The top 50 mutations, the MHC alleles associated with the peptides, and the corresponding changes in activity were identified. Statistical significance was assessed using the Wilcoxon test in Python.

Results

Model performance

Exploring multiple approaches to predict IFNγ release from T-helper cells resulted in the development of eleven models. Among the algorithms tested, the RF consistently outperformed SVM and GBM across all representations (see Supplementary Table 3). The simplest model, built using LBE descriptors, demonstrated the best overall performance (Table 1). The z-scale model and the LBE 15mer, which used the same dataset, showed nearly identical performance. Thus, the additional information provided by the z-scale descriptors did not significantly improve the model's performance. Meanwhile, the ProtBert model resulted in the lowest sensitivity but the highest specificity among all models. This indicates that it was less effective as the other models in capturing the active samples, but better regarding the inactive ones. The standard deviation of the different metrics presented in Table 1 shows that the different methods did not result in notable differences in stability; however, considering all the metrics, the z-scale based model appeared to be the least stable one.

The AL approach was applied to the LBE model to assess whether the model's generalizability could be further improved. After 351 iterations, the model reached the highest mean MCC and was selected as the final model (see Supplementary Fig. 1). The performance across all folds is presented in Supplementary Fig. 2. While the AL model did not surpass the performance of the LBE model, the difference was minimal.

T-cell proliferation dataset performance

The T-cell proliferation dataset was used to assess the model's performance. As expected, the model did not perform as well on the T-cell proliferation dataset as it did on the internal IFNγ release test set. However, it still correctly identified the majority of active samples, as presented in Table 2. Notably, the model achieved a higher sensitivity while maintaining the same precision as in the IFNγ release test set. However, the specificity results indicate that it struggled to accurately predict inactive samples. The confusion matrix in Fig. 3 illustrates the predictions for the T-cell proliferation dataset. Given that 84% of this dataset consisted of active samples, the model successfully classified most of them correctly.

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

Table 1. The table presents the model performances for all representations using a RF algorithm. The values are presented as the mean metric of the 10-fold CV, with the standard deviation indicated in parentheses. The model demonstrating the best overall performance is shown in bold. The representations used are LBE, z-scale descriptors, and ProtBert embedding features. In addition, an AL approach was applied with the LBE representation.

Method	Algorithm	Balanced accuracy	MCC	Precision	Sensitivity	Specificity
LBE	**RF**	**0.78 (0.01)**	**0.53 (0.02)**	**0.88 (0.01)**	**0.78 (0.02)**	**0.77 (0.03)**
LBE 15mer	RF	0.75 (0.01)	0.49 (0.02)	0.85 (0.01)	0.73 (0.02)	0.78 (0.03)
Z-scale	RF	0.76 (0.02)	0.50 (0.04)	0.86 (0.01)	0.74 (0.03)	0.78 (0.03)
AL (LBE)	RF	0.76 (0.02)	0.51 (0.03)	0.86 (0.02)	0.74 (0.03)	0.78 (0.03)
ProtBert	RF	0.76 (0.01)	0.50 (0.03)	0.89 (0.01)	0.72 (0.03)	0.81 (0.03)

Table 2. The result of the T-cell proliferation dataset using the LBE model is presented. The values represent the mean performance, with the standard deviation shown in parentheses.

Method	Balanced accuracy	MCC	Precision	Sensitivity	Specificity
LBE	0.61 (0.02)	0.21 (0.03)	0.88 (0.01)	0.87 (0.01)	0.35 (0.05)

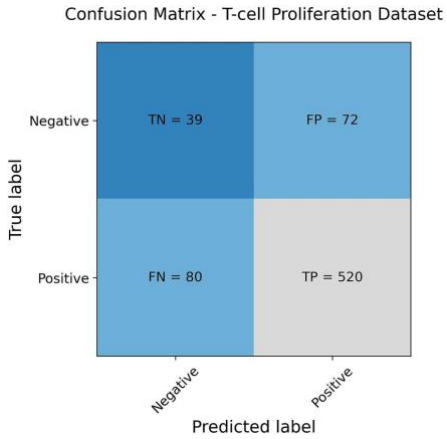


Figure 3. The figure depicts the confusion matrix obtained from the T-cell proliferation dataset, highlighting the high false positive rate. However, it also shows that the model was able to capture most of the active samples.

Feature importance

The LBE 15mer model was used for the feature importance analysis to identify the most influential amino acid positions in the peptides. This analysis was conducted with the 10-fold CV, and the final top five features, determined from all results, are presented in Table 3. The most important position (P) was P3, which is known to be involved in T-cell binding. Two additional positions known to be involved in T-cell binding are P2 and P8, which also ended up among the top five features. The remaining two other positions among the top five features, P14 and P13, are unlikely to be directly involved in either MHC-II binding or T-cell binding. The amino acid frequency was analysed at the top five most important positions, revealing no notable differences between the active and inactive classes (see Fig. 4). The most frequent amino acid in all top five positions was leucine for both the active and inactive classes. Therefore, no notable discrepancies were observed in the amino acid frequency between the two classes.

Table 3. The top five most important features and their rank are presented in the table. The features represent the different positions (P) in the peptide sequence.

Rank	Positions
1	P3
2	P14
3	P2
4	P8
5	P13

Virtual peptide single amino acid mutation

Further insight was acquired by evaluating how individual amino acid variations at a given position in the peptide sequence could impact the prediction. A virtual single amino acid mutation was applied to each position in the peptide sequence, and the LBE 15mer-based model was used to predict the outcomes. The results indicate that predictions were influenced by both the given amino acid and its position. Figure 5 visualizes the ΔER based on the specific mutation in the peptide sequence. The top five positions with the highest ΔER were P2, P3, P8, P13, and P14, as illustrated in Fig. 5A. Four of these positions were also ranked among the top five ones in the feature importance analysis.

A tyrosine mutation at positions P2 and P14 resulted in the highest ΔER of 0.017, followed by valine and tryptophan mutations at P2, each yielding a ΔER of 0.016. In contrast, an asparagine mutation in P9 produced the lowest ΔER observed, as shown in Fig. 5B. These findings highlight the amino acids that play a crucial role at specific positions in the peptide in influencing prediction variability. The same amino acid mutated in different positions can have varying results, as was observed for tyrosine, which obtained a negative ΔER in P6, P7, and P12, but a positive one in P2 and P14.

Furthermore, the results shown in Fig. 5 highlight only the mutated amino acid responsible for the change in prediction, without considering the original amino acid that was mutated. To address this, Supplementary Figs. 3–10 provide detailed information on the count of prediction changes associated with each original and mutated amino acid for specific MHC alleles. This analysis is presented separately for the two classes, focusing

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

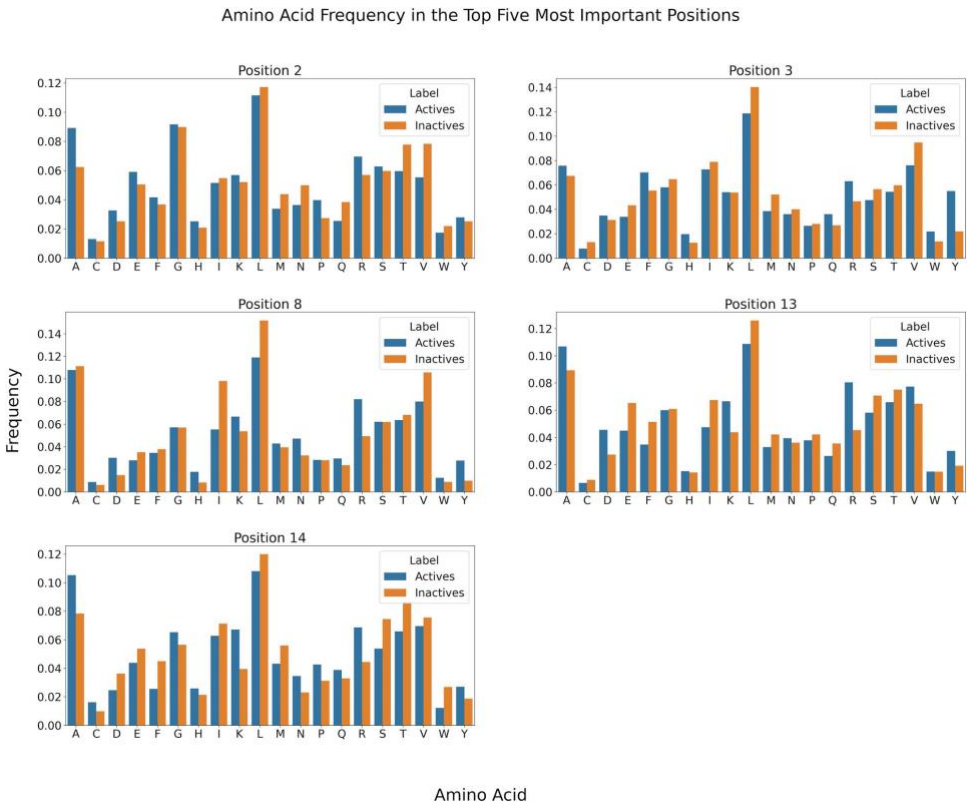


Figure 4. The amino acid frequency in the top five most important positions for the active and inactive classes.

on the five most important positions identified in the feature importance analysis, as well as the mutations that resulted in a statistically significant ΔER at these positions. Table 4 lists the mutated amino acids and their positions that resulted in a P-value < 0.05. Notably, none of the mutated amino acids in P8 resulted in a P-value < 0.05. Some mutations exhibited variable effects depending on the associated MHC allele, meaning they could shift the prediction from active to inactive or vice versa. For instance, when alanine, aspartic acid, and glycine were mutated to tyrosine, the resulting prediction varied between active and inactive depending on the MHC allele linked to the peptide. Specifically, tyrosine, which resulted in the highest ΔER in P2, caused a change in prediction from active to inactive when replacing glycine in peptides associated with MHC alleles HLA-DRB10901 and HLA-DRB10102 (see Supplementary Fig. 3). However, as shown in Supplementary Fig. 7, the same mutation in inactive peptides resulted in a change when the peptide was associated with MHC allele HLA-DRB10802.

Table 4. The mutated amino acids that resulted in a statistically significant change in ΔER in the positions ranked among the top five most important features are presented. The mutated amino acids in P8 did not result in a statistically significant ΔER.

Position	Mutated amino acids
P2	R, Y, W, T, V
P3	K, N, Q
P13	R, N, P, T, S, V, Q
P14	Y, G, W, T, S, V

Discussion

This study focused on building IFNγ release models using information from peptide sequences and MHC-II alleles pseudo-sequences. Furthermore, various descriptors were explored to describe the amino acid sequences. Information-rich descriptors, such as z-scale and ProtBert embedding features, were compared to the most straightforward method, the LBE model. The z-scale descriptors capture physicochemical properties, such as hydrophobicity, steric, and electronic properties. In contrast, the ProtBert model, trained on a large set of protein sequences, identifies complex patterns within amino acid sequences, providing a more biologically meaningful representation. Nonetheless, considering the overall model performance, the models trained on these information-rich descriptors did not outperform the LBE models. One possible explanation for the lack of significant improvement could be due to the increased input dimensionality. Both the z-scale and ProtBert features expand the complexity of

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

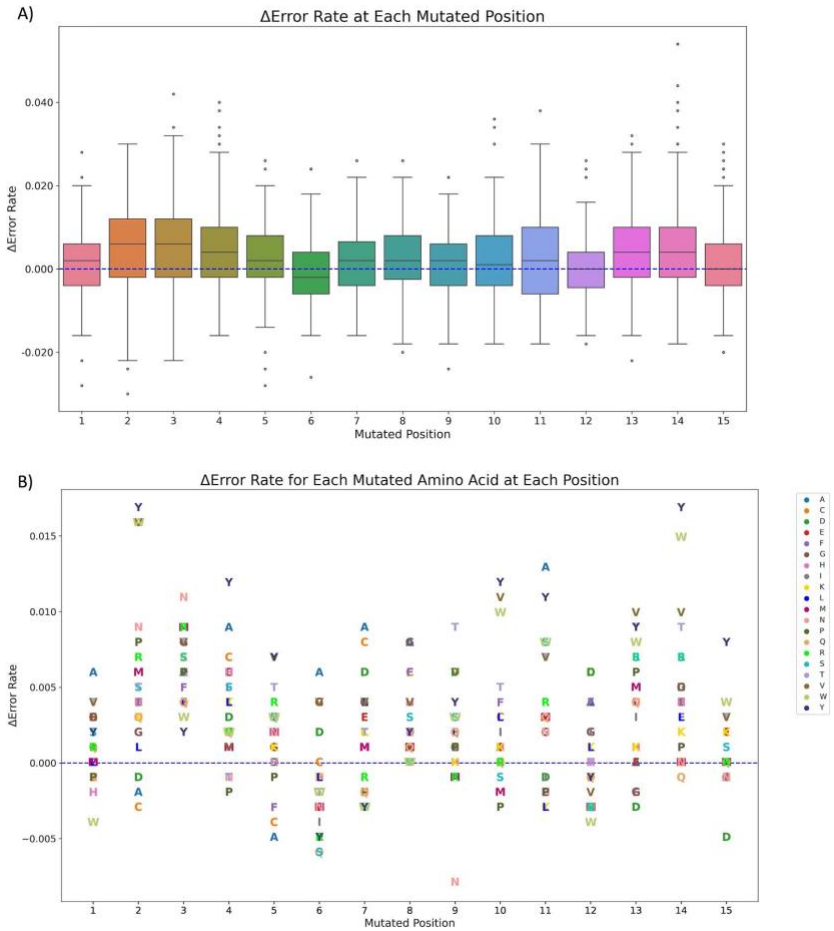


Figure 5. The figure depicts the impact on the prediction after the virtual single amino acid mutation in different positions using the 15mer model. The ΔER was calculated and used as a metric to evaluate the impact of the predictions, where a ΔER value of zero indicates that the prediction was not impacted by the mutation. A) The ΔER for all positions and the median ΔER are depicted in the boxes. B) The ΔER for each mutated amino acid at each position.

the input space, potentially making it more challenging for the model to effectively learn from the information. Another factor that complicates the model learning process is the imbalanced dataset, which can reduce generalizability. To address this, we applied an AL approach to the LBE model, as it demonstrated the best performance. AL has the potential to increase the model's performance or improve generalizability while using fewer data points. This was achieved by selecting the top 10 data points with the highest uncertainty for each iteration and adding them to the training set. However, also this strategy did not improve the model's performance or its ability to correctly predict more inactive samples. As a result, the model remained limited in its capacity to capture the inactive class, which represents the minority class in the dataset. In addition, as shown in Supplementary Fig. 2, the model's performance did not improve with additional iterations beyond 351. The mean MCC value fluctuated within a narrow range of 0.49–0.51, suggesting that the model reached a learning plateau.

Predicting T-cell response is a highly complex task involving multiple steps. However, the models demonstrated sufficient performance for this task. To further assess their applicability, we investigated whether the best-performing model, the LBE model, could predict an activity closely related to IFNγ release. For this purpose, the model was applied to the T-cell proliferation dataset. The prediction resulted in a precision and sensitivity close to 0.90, indicating that the model was very good at capturing active samples. However, it struggled with correctly predicting inactive samples. Given that only 16% of the T-cell proliferation dataset consisted of inactive samples, the evaluation metrics for the minority class may be skewed and may not fully reflect the model's ability to capture inactives. Nonetheless, the model achieved a balanced accuracy of 0.61. Considering that the T-cell proliferation dataset

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

is highly imbalanced and represents a slightly different endpoint than the dataset used to build the model, this result suggests that the model was able to capture meaningful information related to a T-cell activity and generalized to a certain extent to the related but distinct dataset.

To further examine and better understand which information was captured and learned by the model, a feature importance analysis was performed. Positions P2, P3, and P8 were identified among the top five most important features, aligning with positions known to be involved in T-cell binding [9]. Additionally, positions P13 and P14 were ranked among the top five positions, despite not being directly involved in MHC-II or T-cell binding. However, these positions can still impact the affinity for the MHC-II molecule as well as the stability of the pMHC-II complex, as demonstrated in various studies [25–27]. Therefore, the specific amino acids at these positions affect the stability of the complex, which in turn influences the outcome of the T-cell response. These findings suggest that the model successfully captured biologically relevant information. To further explore potential differences between active and inactive samples, the amino acid frequency at the top five positions was analysed. No significant discrepancies were observed, indicating that factors beyond the presence of a single amino acid at a given position contribute to the observed activity. Subsequently, we conducted a virtual single amino acid mutation at every position to assess the impact on the model's predictions. Four of the positions with the highest feature importance also exhibited the highest median ΔER values, suggesting that mutations at these positions were more likely to impact the prediction. However, the results also revealed that the effect varied depending on the specific amino acid, indicating that prediction changes were not solely determined by the position but also influenced by the amino acid itself.

Sant'Angelo et al. showed that an alanine mutation in P2 led to an increased T-cell response [9]. In our analysis, five amino acids at position P2 showed a statistically significant ΔER, indicating that mutations at this position influenced the model's prediction. This suggests that the model may associate these mutations with a potential change in T-cell response. Furthermore, the results showed that, depending on the MHC allele, the same mutation could result in an active or inactive prediction. Additionally, in some cases, mutations involving two specific amino acids produced opposite activity outcomes. For instance, mutating threonine to tyrosine in peptides associated with the MHC allele HLA-DRB10301 resulted in a shift from active to inactive. Conversely, the reverse mutation (tyrosine to threonine) for the same MHC allele resulted in a prediction change from inactive to active. These results, combined with the lack of pronounced ΔER values, suggest a synergistic interplay among multiple positions and distinct amino acids. This finding implies that the model was integrating additional features by identifying patterns and making decisions based on a more comprehensive evaluation.

As previously established, T-cell response is a highly complex process with various factors contributing. Consequently, our approach had certain limitations. One aspect was that the data was assessed by different assay formats, which was not explicitly accounted for when building the model. This also applies to other assay-specific details that were not considered. Additionally, the T-cell proliferation dataset was highly imbalanced, which may have affected the model's ability to accurately predict inactive samples. However, this limitation was also due to the shortcomings in publicly available databases containing T-cell assay data. Furthermore, this study utilized as many MHC alleles as possible to provide users with a broader applicability. However, it is important to note that the frequency of MHC alleles varies across populations, with some alleles being more prevalent than others. This imbalance was also reflected in our dataset and may have influenced the prediction accuracy for less common MHC alleles.

According to the authors of ProtBert, fine-tuning the ProtBert model has shown to provide higher accuracy in certain tasks compared to using the embedding features. This approach could potentially enhance our model's performance as well. Additionally, overparametrized models have been shown to improve model generalizability [28, 29], suggesting that more complex BERT models may offer a better outcome. Hence, future work could focus on fine-tuning large models, as well as utilizing the more complex BERT models. Furthermore, incorporating additional assay-related information into the model may enhance predictive accuracy. Moreover, given that our results suggest a synergistic interplay among multiple factors, a subsequent next step could be to investigate these interactions further.

Conclusion

To explore various approaches for building IFNγ release models, different representations were used to describe peptide sequences and MHC-II alleles pseudo-sequences. In addition, feature importance and virtual single amino acid mutations were performed to gain insights into the model's decision-making process and to improve model interpretability. The model was tested for its generalizability beyond IFNγ release by predicting T-cell proliferation, a process closely associated with IFNγ release. The findings from this study highlight the model's applicability and its potential to support the evaluation of a T-cell activity closely related to IFNγ release.

Key Points

- Investigated different approaches for developing IFNγ release models.
- Evaluated the model's generalizability to a T-cell activity closely associated with IFNγ release.
- Conducted feature importance analysis and virtual amino acid mutations to enhance model interpretability.

Supplementary data

Supplementary data is available at *Briefings in Bioinformatics* online.

Conflict of interest: None declared.

Funding

This research was funded in whole or in part by the Austrian Science Fund (FWF) AW012321. For open access purposes, the author has applied a CC BY public copyright license to any author accepted manuscript version arising from this submission. The study was also funded by Sanofi.

Data availability

Data and source code utilized in this work are available at: https://github.com/PharminfoVienna/IFNg_models.

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

References

1. Dimitrov DS. Therapeutic Proteins. In: Voynov V, Caravella JA, eds. *Therapeutic Proteins. Methods in Molecular Biology*. Humana Press; 2012;**899**:1–26. https://doi.org/10.1007/978-1-61779-921-1_1
2. Chames P, Van Regenmortel, Weiss E. *et al.* Therapeutic antibodies: successes, limitations and hopes for the future: therapeutic antibodies: an update. *Br J Pharmacol* 2009;**157**:220–33. https://doi.org/10.1111/j.1476-5381.2009.00190.x
3. Muromonab-CD3. *in LiverTox: Clinical and Research Information on Drug-Induced Liver Injury*. Bethesda (MD): National Institute of Diabetes and Digestive and Kidney Diseases, 2012. Accessed: Jul. 25, 2024. [Online]. Available: http://www.ncbi.nlm.nih.gov/books/NBK548590/.
4. Shahbandeh M. Protein products market - statistics & facts. Statista. Accessed: Jul. 25, 2024. [Online]. Available: https://www.statista.com/topics/4232/protein-market/
5. Rammensee H-G, Bachmann J, Emmerich NPN. *et al.* SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 1999;**50**:213–9. https://doi.org/10.1007/s002510050595
6. Chicz RM, Urban RG, Lane WS. *et al.* Predominant naturally processed peptides bound to HLA-DR1 are derived from MHC-related molecules and are heterogeneous in size. *Nature* 1992;**358**:764–8. https://doi.org/10.1038/358764a0
7. Ochoa R, Lunardelli VAS, Rosa DS. *et al.* Multiple-allele MHC class II epitope engineering by a molecular dynamics-based evolution protocol. *Front Immunol* 2022;**13**. https://doi.org/10.3389/fimmu.2022.862851
8. Ochoa R, Laskowski RA, Thornton JM. *et al.* Impact of structural observables from simulations to predict the effect of single-point mutations in MHC class II peptide binders. *Front Mol Biosci* 2021;**8**. https://doi.org/10.3389/fmolb.2021.636562
9. Sant'Angelo DB, Robinson E, Janeway CA Jr. *et al.* Recognition of core and flanking amino acids of MHC class II-bound peptides by the T cell receptor. *Eur J Immunol* 2002;**32**:2510–20. https://doi.org/10.1002/1521-4141(200209)32:9<2510::AID-IMMU2510>3.0.CO;2-Q
10. Berger A. Th1 and Th2 responses: what are they? *BMJ* 2000;**321**:424. https://doi.org/10.1136/bmj.321.7258.424
11. Zhao H, Wu L, Yan G. *et al.* Inflammation and tumor progression: signaling pathways and targeted intervention. *Sig Transduct Target Ther* 2021;**6**:263. https://doi.org/10.1038/s41392-021-00658-5
12. Nonclinical Evaluation of the Immunotoxic Potential of Pharmaceuticals. U.S. Department of Health and Human Services, Food and Drug Administration, Center for Drug Evaluation and Research (CDER); 2023.
13. Rombach-Riegraf V, Karle AC, Wolf B. *et al.* Aggregation of human recombinant monoclonal antibodies influences the capacity of dendritic cells to stimulate adaptive T-cell responses in vitro. *PloS One* 2014;**9**:e86322. https://doi.org/10.1371/journal.pone.0086322
14. De Wolf C, Van De Bovenkamp M, Hoefnagel M. Regulatory perspective on in vitro potency assays for human T cells used in anti-tumor immunotherapy. *Cytotherapy* 2018;**20**:601–22. https://doi.org/10.1016/j.jcyt.2018.01.011
15. Binayke A, Zaheer A, Vishwakarma S. *et al.* A quest for universal anti-SARS-CoV-2 T cell assay: systematic review, meta-analysis, and experimental validation. *npj Vaccines* 2024;**9**:3. https://doi.org/10.1038/s41541-023-00794-9
16. Tung C-W, Ziehm M, Kämper A. *et al.* POPISK: T-cell reactivity prediction using support vector machines and string kernels. *BMC Bioinformatics* 2011;**12**:446. https://doi.org/10.1186/1471-2105-12-446
17. Reynisson B, Barra C, Kaabinejadian S. *et al.* Improved prediction of MHC II antigen presentation through integration and motif deconvolution of mass spectrometry MHC eluted ligand data. *J Proteome Res* 2020;**19**:2304–15. https://doi.org/10.1021/acs.jproteome.9b00874
18. Cheng J, Bendjama K, Rittner K. *et al.* BERTMHC: improved MHC–peptide class II interaction prediction with transformer and multiple instance learning. *Bioinformatics* 2021;**37**:4172–9. https://doi.org/10.1093/bioinformatics/btab422
19. Gasser HC, Bedran G, Ren B, Goodlett D, Alfaro J, Rajan A. Interpreting BERT architecture predictions for peptide presentation by MHC class I proteins. Published online November 13, 2021. https://doi.org/10.48550/arXiv.2111.07137
20. Sandberg M, Eriksson L, Jonsson J. *et al.* New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. *J Med Chem* 1998;**41**:2481–91. https://doi.org/10.1021/jm9700575
21. Dhanda SK, Vir P, Raghava GP. Designing of interferon-gamma inducing MHC class-II binders. *Biol Direct* 2013;**8**:30. https://doi.org/10.1186/1745-6150-8-30
22. Karosiene E, Rasmussen M, Blicher T. *et al.* NetMHCIIpan-3.0, a common pan-specific MHC class II prediction method including all three human MHC class II isotypes, HLA-DR, HLA-DP and HLA-DQ. *Immunogenetics* 2013;**65**:711–24. https://doi.org/10.1007/s00251-013-0720-y
23. Elnaggar A, Heinzinger M, Dallago C. *et al.* ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381
24. S. Ertekin, J. Huang, L. Bottou, and L. Giles, 'Learning on the border: active learning in imbalanced data classification', in *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, Lisbon Portugal: ACM, 2007, pp. 127–36. doi: https://doi.org/10.1145/1321440.1321461
25. Chang ST, Ghosh D, Kirschner DE. *et al.* Peptide length-based prediction of peptide–MHC class II binding. *Bioinformatics* 2006;**22**:2761–7. https://doi.org/10.1093/bioinformatics/btl479
26. O'Brien C, Flower DR, Feighery C. Peptide length significantly influences in vitro affinity for MHC class II molecules. *Immunome Res* 2008;**4**:6. https://doi.org/10.1186/1745-7580-4-6
27. Nelson CA, Petzold SJ, Unanue ER. Identification of two distinct properties of class II major histocompatibility complex-associated peptides. *Proc Natl Acad Sci USA* 1993;**90**:1227–31. https://doi.org/10.1073/pnas.90.4.1227
28. Belkin M, Hsu D, Ma S. *et al.* Reconciling modern machine learning practice and the bias-variance trade-off. *Proc Natl Acad Sci U S A* 2019;**116**:15849–54. https://doi.org/10.1073/pnas.1903070116
29. Simon JB, Karkada D, Ghosh N, Belkin M. More is Better in Modern Machine Learning: when Infinite Overparameterization is Optimal and Overfitting is Obligatory. Published online May 15, 2024. https://doi.org/10.48550/arXiv.2311.14646

© The Author(s) 2025. Published by Oxford University Press. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.
Briefings in Bioinformatics, 2025, **26(2)**, bbaf101
https://doi.org/10.1093/bib/bbaf101
Problem Solving Protocol

Downloaded from https://academic.oup.com/bib/article/26/2/bbaf101/8069414 by guest on 17 June 2025

Chapter 5: Concluding Discussion and Outlook

The main objective of this thesis was to investigate the potential of computational approaches to predict and understand the toxicity of biological drugs through three studies: analyzing side effect profiles, evaluating the predictiveness of animal models, and predicting immunogenicity using machine learning. The studies used data from CLARITY, PharmaPendium, and IEDB and aimed to uncover both methodological insights and practical applications.

The first study centered on the development of an interactive, open-source dashboard to analyze and visualize side effect data from multiple sources. By integrating data from CLARITY and PharmaPendium and incorporating MedDRA's hierarchical structure, the dashboard enables users to explore side effects across three perspectives: "Drug", "Target" and "Adverse Event". It also supports analysis across datasets ranging from preclinical (including species information) to clinical and post-marketing data. The key strength of the dashboard lies in its ability to combine data from different sources and visualize them consistently using MedDRA's hierarchy. This allows users to move between — e.g. from specific PTs to broader high-level terms to gain a comprehensive view of side effect patterns. This proved especially valuable in selecting use cases for further study, such as neutropenias, which include multiple related PTs that can be better grouped under a common HLT, neutropenias. However, the dashboard has limitations. It requires users to provide side effect data using PTs, and they must supply their datasets and have access to a MedDRA license to map the PTs to their respective higher-level terms. Despite these constraints, the dashboard offers a scalable and flexible tool for exploring side effect data across different stages of drug development.

The second study investigated the side effect landscape for therapeutic mAbs and the difference between side effects reported as AEs and ADRs. In addition, the study also aimed to assess the ability of preclinical data to predict side effects in humans through an animal-to-human translation. By creating separate toxicity profiles for side effects categorized as either AEs or ADRs for each therapeutic mAb and clustering them, we were able to visualize and analyze their side effect landscape. This enabled an evaluation of whether AE- and ADR-derived profiles led to different groupings of drugs. Additionally, by incorporating both AE and ADR data into the animal-to-human translation, we assessed whether animal models could predict neutropenia and whether predictive value differed depending on the reporting type, AE or ADR. In general, mAbs clustered by their therapeutic area, as defined by the

Anatomical Therapeutic Chemical (ATC) level 2 code, and their target class. By differentiating the side effect landscape between AEs and ADRs, we could assess whether any discrepancies emerged. The ADR landscape was more clearly defined by therapeutic area compared to the AE landscape, suggesting that the ADR landscape of a mAb may potentially be estimated based on its therapeutic area and target class. This analysis was based on PT level, though the neutropenia use case applied the HLT level, which aggregates related PTs.

For the neutropenias use case, ATC level 2 code appeared predictive when neutropenia was reported as an AE but not as an ADR. This discrepancy may reflect differences in aggregation; had PT level data been used, results might have varied, meaning that if a specific type of neutropenia at the PT level had been investigated instead, the results might have differed. All anti-interleukin-6/interleukin-6-receptor (IL-6/IL-6R) mAbs were identified using ADR data at the HLT level — these drugs are known to have downstream effects that can cause neutropenia, although the exact mechanism remains unclear [24]. Thus, ADR data at the HLT level appear more suited to identify mechanistically related drugs, providing added biological insight. The findings from human data were therefore compared to those from animal data.

The animal-to-human translation revealed a lack of concordance between datasets, indicating that animal data are not reliable predictors of HLT level neutropenias. The analysis included all species, and only one target class overlapped between animal and human datasets. The findings suggests that animal data provided limited predictive value. Notably, anti-IL-6/IL-6R mAbs were entirely absent in the animal dataset. This highlights species differences and supports the view that therapeutic area and target class provide more robust predictors than animal data.

With the continued expansion of biological drugs such as mAbs, leveraging existing structured side effect data is essential to understand the factors associated with their side effects. Studies 1 and 2 focused on understanding this structure and evaluating the predictive power of therapeutic information. The limited value of animal data underscores the importance of integrating human-derived information for early toxicity assessment of therapeutic mAbs.

Some side effects, however, are induced by the activation of the immune response [6], making immunogenicity a critical aspect of biological drug development for both safety and efficacy. Therefore, the third and final study applied machine learning to predict immunogenicity based on peptide sequences and MHC class II pseudo sequences. Unlike existing methods focused on MHC binding, this study aimed to evaluate immunogenicity further downstream in the immune response. The IFNγ release dataset from IEDB, representing the largest available endpoint, was selected as the starting point.

Due to the dataset's limited size of 7266 data points and class imbalance, the main goal was to identify which descriptors and methods performed best and to understand whether the model features were relevant to immunogenicity. To ensure robust model evaluation, stratified 10-fold cross-validation and hyperparameter optimization were applied throughout. The best results were obtained using the simplest descriptor, the letter-based encoding (LBE) combined with the RF classifier.

To address the class imbalance, several strategies were employed. A stratified train-test split ensured consistent class distribution across sets, maximizing learning from the minority class while preserving enough data for evaluation. Additionally, adjusting the classification threshold helped reduce bias towards the majority class — in this case, the active class, which otherwise risked overfitting. A probability threshold of 0.65 resulted in a more balanced trade-off between sensitivity and specificity.

Depending on the application, certain metrics take precedence — particularly in risk assessment, where a higher number of FP may be preferable to FN, as toxicity pose a greater threat and a cautious approach is warranted. However, since immunogenicity can both be desired (e.g. vaccines) and undesired the MCC was used as the main evaluation metric due to its balanced consideration of both classes.

Active learning was also implemented and tested as well as using ProtBert embedding vectors, due to the limited size of the dataset and the class imbalance. However, it did not yield better results; while for the active learning, fewer data points were required to reach a comparable level, overall performance was slightly reduced.

For the method that yielded the best overall performance, a feature importance analysis was conducted. The findings showed that the model had learned biologically relevant patterns, with several of the most influential features corresponding to known T-cell binding positions. A subsequent analysis, involving virtual amino acid mutation, demonstrated that different amino acids in different positions influenced whether the model predicted a data point as active or inactive. Moreover, the analysis suggested a synergistic interplay, where specific amino acids in specific positions collectively determined the prediction outcome.

The three studies presented in this thesis addressed the side effects of biological drugs. Although biological drugs are becoming increasingly sought after, the number of approved biological drugs, such as therapeutic mAbs, remains relatively low compared to small molecule drugs. As a result, data scarcity poses a significant challenge for model development and risk assessment.

Biological drugs often affect the immune system, which is influenced by numerous factors such as genetic, pre-existing conditions, and other variables that are difficult to measure. This adds further complexity to data interpretation and model development. A key limitation of this work lies in the limited availability of relevant data and the challenges this entails. However, as more biologicals are approved and computational methods continue to evolve, our understanding of how to collect and structure data to support the analysis and prediction of side effects of biological drugs will improve. For instance, in Study 2, it would be both valuable and interesting to analyze side effects in the context of the specific disease for which drug was used. If such indications were consistently annotated in databases, it would enable a more nuanced differentiation of side effects — not only based on high-level therapeutic categories but also on actual clinical use. Another limitation, in Study 3 is that currently, databases such as IEDB include assays based on viral antigens, which differ from antigens derived from therapeutic mAbs. Expanding data collecting specifically for therapeutic mAbs would help improve existing prediction methods and make them more applicable to these types of molecules.

Data availability was one of the key limitations in this thesis. However, there are several ways in which the work presented in this thesis could be expanded and improved. For instance, in Study 2, the analysis could be extended to include higher MedDRA levels to explore whether differences between AEs and ADRs persist across these levels, and whether

the observed patterns change. It would also be valuable to investigate whether the findings from the animal-to-human translation remain consistent when using different levels of side effects classification, and whether this applies broadly across various side effects. As such, constructing a side effect landscape based on animal data could be an interesting next step for this study. For Study 3, future work could involve incorporating additional information into the IFNγ release model and explore other endpoints. Additionally, it would be valuable to further explore the synergistic interplay identified in the current dataset.

The studies in this thesis aimed to investigate and predict the toxicity of biological drugs using data from three different databases. These studies have provided valuable insights into the factors influencing the toxicity of biological drugs. Therefore, the tools developed and the results obtained can be useful in facilitating ongoing research in this area and in improving the understanding of the toxicity of biological drugs based on the data used in this work, ultimately supporting safer drug development and monitoring.

Bibliography

[1] O. of the Commissioner, 'The Drug Development Process', FDA. Accessed: May 22, 2025. [Online]. Available: https://www.fda.gov/patients/learn-about-drug-and-device-approvals/drug-development-process

[2] H. Prior, F. Sewell, and J. Stewart, 'Overview of 3Rs opportunities in drug discovery and development using non-human primates', *Drug Discovery Today: Disease Models*, vol. 23, pp. 11–16, 2017, doi: 10.1016/j.ddmod.2017.11.005.

[3] 'Regulatory Knowledge Guide Biological Products'.

[4] R. P. Evens, 'Pharma Success in Product Development—Does Biotechnology Change the Paradigm in Product Development and Attrition', *AAPS J*, vol. 18, no. 1, pp. 281–285, Jan. 2016, doi: 10.1208/s12248-015-9833-6.

[5] D. Shah, B. Soper, and L. Shopland, 'Cytokine release syndrome and cancer immunotherapies – historical challenges and promising futures', *Front. Immunol.*, vol. 14, p. 1190379, May 2023, doi: 10.3389/fimmu.2023.1190379.

[6] E. S. Björnsson *et al.*, 'Risk of Drug-Induced Liver Injury From Tumor Necrosis Factor Antagonists', *Clinical Gastroenterology and Hepatology*, vol. 13, no. 3, pp. 602–608, Mar. 2015, doi: 10.1016/j.cgh.2014.07.062.

[7] J. Krämer and H. Wiendl, 'What Have Failed, Interrupted, and Withdrawn Antibody Therapies in Multiple Sclerosis Taught Us?', *Neurotherapeutics*, vol. 19, no. 3, pp. 785–807, Apr. 2022, doi: 10.1007/s13311-022-01246-3.

[8] C. De Wolf, M. Van De Bovenkamp, and M. Hoefnagel, 'Regulatory perspective on in vitro potency assays for human T cells used in anti-tumor immunotherapy', *Cytotherapy*, vol. 20, no. 5, pp. 601–622, May 2018, doi: 10.1016/j.jcyt.2018.01.011.

[9] C. Marks, A. M. Hummer, M. Chin, and C. M. Deane, 'Humanization of antibodies using a machine learning approach on large-scale repertoire data', *Bioinformatics*, vol. 37, no. 22, pp. 4041–4047, Nov. 2021, doi: 10.1093/bioinformatics/btab434.

[10] B. Reynisson, C. Barra, S. Kaabinejadian, W. H. Hildebrand, B. Peters, and M. Nielsen, 'Improved Prediction of MHC II Antigen Presentation through Integration and Motif Deconvolution of Mass Spectrometry MHC Eluted Ligand Data', *J. Proteome Res.*, vol. 19, no. 6, pp. 2304–2315, Jun. 2020, doi: 10.1021/acs.jproteome.9b00874.

[11] A. E. Mattei *et al.*, 'In silico methods for immunogenicity risk assessment and human homology screening for therapeutic antibodies', *mAbs*, vol. 16, no. 1, p. 2333729,

Dec. 2024, doi: 10.1080/19420862.2024.2333729.
[12] ‘Adverse event | European Medicines Agency’. Accessed: Jul. 16, 2024. [Online]. Available: https://www.ema.europa.eu/en/glossary/adverse-event
[13] ‘Adverse drug reaction | European Medicines Agency’. Accessed: Jul. 29, 2024. [Online]. Available: https://www.ema.europa.eu/en/glossary/adverse-drug-reaction
[14] J. Yang *et al.*, ‘Brief introduction of medical database and data mining technology in big data era’, *J Evidence Based Medicine*, vol. 13, no. 1, pp. 57–69, Feb. 2020, doi: 10.1111/jebm.12373.
[15] ‘Welcome to MedDRA | MedDRA’. Accessed: Mar. 01, 2024. [Online]. Available: https://www.meddra.org/
[16] ‘PharmaPendium | Drug safety, efficacy & DMPK data | Elsevier’, www.elsevier.com. Accessed: Jun. 12, 2023. [Online]. Available: https://www.elsevier.com/products/pharmapendium
[17] S. J. W. Evans, P. C. Waller, and S. Davis, ‘Use of proportional reporting ratios (PRRs) for signal generation from spontaneous adverse drug reaction reports’, *Pharmacoepidemiology and Drug*, vol. 10, no. 6, pp. 483–486, Oct. 2001, doi: 10.1002/pds.677.
[18] ‘What Is Random Forest? | IBM’. Accessed: May 02, 2025. [Online]. Available: https://www.ibm.com/think/topics/random-forest
[19] ‘What Is Support Vector Machine? | IBM’. Accessed: May 02, 2025. [Online]. Available: https://www.ibm.com/think/topics/support-vector-machine
[20] ‘What is Gradient Boosting? | IBM’. Accessed: May 02, 2025. [Online]. Available: https://www.ibm.com/topics/gradient-boosting
[21] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjöström, and S. Wold, ‘New Chemical Descriptors Relevant for the Design of Biologically Active Peptides. A Multivariate Characterization of 87 Amino Acids’, *J. Med. Chem.*, vol. 41, no. 14, pp. 2481–2491, Jul. 1998, doi: 10.1021/jm9700575.
[22] U. Abu Shehab and G. F. Ecker, ‘To what extent can we extrapolate Proteochemometric models: A case study for the SLC6 transporter family’, Jan. 28, 2025. doi: 10.26434/chemrxiv-2025-sbw2l-v2.
[23] A. Elnaggar *et al.*, ‘ProtTrans: Towards Cracking the Language of Life’s Code Through Self-Supervised Learning’, Jul. 12, 2020. doi: 10.1101/2020.07.12.199554.
[24] T. S. Mitchell, R. J. Moots, and H. L. Wright, ‘Janus kinase inhibitors prevent migration of rheumatoid arthritis neutrophils towards interleukin-8, but do not inhibit priming

of the respiratory burst or reactive oxygen species production', *Clinical and Experimental Immunology*, vol. 189, no. 2, pp. 250–258, Jul. 2017, doi: 10.1111/cei.12970.

Abstract

Biological drugs are large and complex molecules that have successfully treated diseases once considered challenging. With the continued growth of biological drugs, the need to understand their side effects and to develop computational tools to better assess their toxicity is increasing. The main aim of this thesis was to explore the toxicity prediction of biological drugs by utilizing data from available databases, analyzing the data, and investigating different ways to describe these macromolecules for machine learning. The thesis consists of three studies, all of which focus on the toxicity of biological drugs.

Study 1 presents an interactive dashboard that utilizes side effect data based on Preferred Terms (PTs) from the MedDRA terminology and maps it to the other hierarchical terms. This enables the user to analyze the data across four different levels for preclinical, clinical, and post-marketing data.

Study 2 investigates the side effect landscape based on side effects reported as adverse events (AEs) and adverse drug reactions (ADRs) for therapeutic mAbs. Additionally, Study 2 focused on neutropenia as a use case, highlighting the lack of concordance between animal models and human outcomes in predicting neutropenia. The findings also showed that, by simply using the therapeutical information of the drug, neutropenia reported as an AE was more predictable than when using animal data.

The final study, Study 3, explores the prediction of immunogenicity by testing different representations and machine learning methods. The result showed that the simplest representation format, combined with a random forest algorithm, resulted in the best overall performance.

Zusammenfassung

Biologische Arzneimittel sind große und komplexe Moleküle, die erfolgreich Krankheiten behandelt haben, die früher als schwer therapierbar galten. Mit dem anhaltenden Wachstum biologischer Arzneimittel steigt der Bedarf, ihre Nebenwirkungen besser zu verstehen und rechnergestützte Werkzeuge zur Bewertung ihrer Toxizität zu entwickeln. Ziel dieser Arbeit war es, die Vorhersage der Toxizität biologischer Arzneimittel zu untersuchen, indem Daten aus verfügbaren Datenbanken genutzt, analysiert und verschiedene Methoden zur Beschreibung dieser Makromoleküle für den Einsatz im maschinellen Lernen getestet wurden. Die Arbeit umfasst drei Studien, die sich alle auf die Toxizität biologischer Arzneimittel konzentrieren.

Studie 1 präsentiert ein interaktives Dashboard, das Nebenwirkungsdaten basierend auf den „Preferred Terms“ der MedDRA-Terminologie verwendet und diese auf die anderen hierarchischen Ebenen abbildet. Dies ermöglicht es den Nutzer:innen, die Daten auf vier verschiedenen Ebenen – präklinisch, klinisch und nach Markteinführung – zu analysieren.

Studie 2 untersucht die Landschaft der Nebenwirkungen, basierend auf Nebenwirkungen, die als unerwünschte Ereignisse (Adverse Events, AEs) und unerwünschte Arzneimittelwirkungen (Adverse Drug Reactions, ADRs) bei therapeutischen monoklonalen Antikörpern (mAbs) gemeldet wurden. Zusätzlich wurde Neutropenie als Anwendungsfall betrachtet, um die mangelnde Übereinstimmung zwischen Tiermodellen und menschlichen Ergebnissen bei der Vorhersage von Neutropenie aufzuzeigen. Die Ergebnisse zeigten außerdem, dass Neutropenie, wenn sie als AE gemeldet wurde, allein durch die therapeutischen Informationen des Arzneimittels besser vorhergesagt werden konnte als mit Tierdaten.

Die letzte Studie, Studie 3, untersucht die Vorhersage der Immunogenität durch den Einsatz verschiedener Repräsentationsformen und Methoden des maschinellen Lernens. Die Ergebnisse zeigten, dass das einfachste Repräsentationsformat in Kombination mit einem Random-Forest-Algorithmus die beste Gesamtleistung erzielte.