University of Brasilia

Institute of Exact Sciences
Department of Computer Science

Improving linear B-cell epitope prediction by transfer-learning from higher to lower taxonomic levels

Lindeberg Pessoa Leite

Document presented for qualifying examination of the Ph.D. Program in Computer Science

Supervisor
Prof. Dr. Teófilo Emidio de Campos

Co-Supervisor
Prof. Dr. Felipe Campelo França Pinto

Brasilia
2023

University of Brasilia

Institute of Exact Sciences
Department of Computer Science

Improving linear B-cell epitope prediction by transfer-learning from higher to lower taxonomic levels

Lindeberg Pessoa Leite

Document presented for qualifying examination of the Ph.D. Program in Computer Science

Prof. Dr. Teófilo Emidio de Campos (Supervisor)
CIC/UnB

Prof. Dr. Felipe Campelo França Pinto (Co-Supervisor)
Aston University

Prof. Dr. Luis Paulo Faina Garcia
CIC/UnB

Prof. Dr. Cristiano Leite de Castro
UFMG

Dr. João Luís Reis Cunha
University of York

Prof. Dr. Ricardo Pezzuol Jacobi
Computer Science Graduate Program Coordinator

Brasilia, November 27, 2023

Abstract

Identification of linear B-cell epitopes (LBCEs) plays a key role in the development of diagnostic tests and vaccines against infectious diseases. However, experimental methods used to determine LBCEs are costly and time-consuming. This has motivated the development of computational methods for the rapid identification of LBCEs based on protein sequence data. To date, multiple machine learning approaches have been developed to address this task. These methods rely on having access to a sufficient amount of epitope data either for training generalist predictive models - which may not generalise well to specific pathogens - or to develop organism-specific predictors, which may suffer from data scarcity, particularly for less studied pathogens. These methods face even greater difficulties when dealing with emerging pathogens due to the lack of samples in current data bases. This thesis investigates the potential of improving the performance of the identification of LBCEs by applying transfer-learning from higher to lower taxonomic levels using taxon-specific pre-trained models. Furthermore, the objective of this research is to establish a comprehensive methodology that integrates evolutionary, physicochemical, and structural attributes of amino acids to enhance the overall feature representation. We observed that by transferring the learned features from specific organisms that are evolutionarily more closely related, the resulting models achieve better performance in predicting linear B-cell epitopes. This leads to increased performance in comparison to state-of-the-art methods for LBCE prediction in terms of AUC, F1, and MCC scores.

Keywords: Linear B-cell epitopes, Taxonomy-aware modelling, Transfer-learning.

Resumo Expandido

A identificação de epítopos lineares de células B (LBCEs) desempenha um papel fundamental no desenvolvimento de testes diagnósticos e vacinas contra doenças infecciosas. No entanto, os métodos experimentais usados para determinar LBCEs são caros e demorados. Isso motivou o desenvolvimento de métodos computacionais para a rápida identificação de LBCEs com base em dados de sequências de proteínas. Até o momento, várias abordagens de aprendizado de máquina foram desenvolvidas para lidar com esta tarefa. Esses métodos dependem do acesso a uma quantidade suficiente de dados de epítopos para treinar modelos preditivos generalistas - que podem não generalizar bem para patógenos específicos - ou para desenvolver preditores específicos de organismos, que podem sofrer com a escassez de dados, especialmente para patógenos menos estudados. Esses métodos enfrentam dificuldades ainda maiores ao lidar com patógenos emergentes, devido à falta de amostras nos bancos de dados atuais. Esta tese investiga o potencial de melhorar o desempenho da identificação de LBCEs aplicando transferência de aprendizado de níveis taxonômicos mais altos para mais baixos, usando modelos pré-treinados específicos de táxons. Além disso, o objetivo desta pesquisa é estabelecer uma metodologia abrangente que integre atributos evolutivos, físico-químicos e estruturais de aminoácidos para aprimorar a representação geral de características. Observamos que, ao transferir as características aprendidas de organismos específicos que são evolutivamente mais próximos, os modelos resultantes alcançam melhor desempenhos na predição de epítopos lineares de células B. Isso leva a um aumento no desempenho em comparação com os métodos de *state-of-the-art* para predição de LBCE em termos de métricas de AUC, F1 e MCC.

Palavras-chave: Epítopos lineares de células B, Modelagem informada por taxonomia, Aprendizado de transferência.

Contents

Chapter 1

Introduction

When a living organism is exposed to pathogen, such as viruses or bacteria, the immune system's B-cells[1] recognize the antigens[2] of the pathogen through their B-cell receptors[3] and generate specific antibodies in response. An antigenic determinant, or B-cell epitope (BCE), is the specific region of an antigen that binds to an immune cell receptor. In the case of protein antigens, these epitopes can either be a short amino acid sequence found within the protein or a cluster of atoms located on the protein surface (Ponomarenko and Regenmortel, 2009).

Identification of linear B-cell epitopes (LBCEs) refers to the task of predicting whether a given contiguous amino acid sequence within a protein corresponds to a B-cell epitope or not. LBCEs play a key role in the development of diagnostic tests and vaccines against infectious diseases (Ashford et al., 2021). However, experimental methods used to determine LBCEs are generally costly and time-consuming. This has motivated the development of computational methods for the rapid identification of LBCEs based on protein sequence data. To date, multiple machine learning approaches have been developed to address this task. These methods rely on having access to a sufficient amount of epitope data for training the models to predict linear B-cell epitopes for a given target organism. However, the availability of epitope data remains a significant challenge, particularly when trying to develop or optimise predictors specifically for less-studied organisms or novel pathogens, where the epitope data is limited or nonexistent (Moris et al., 2020). This lack of diverse and representative training data can lead to biased and unreliable predictions, which can in turn affect the efficacy of the method in prioritising targets for downstream experimental assessment.

[1] B cells are white blood cells that produce antibodies to combat foreign substances such as pathogens.

[2] An antigen is a foreign substance that triggers an immune response(Merriam-Webster, 2023b)

[3] B-cell receptors (BCRs) are specialized proteins found on the surface of B cells, which are a type of white blood cell involved in the immune response.

Epitope prediction has been demonstrated to provide several benefits in recent research on epitopes for diagnostic applications. Firstly, it enhances test specificity by identifying highly antigenic regions, reducing false-positive results and ensuring more accurate identification of affected individuals. Secondly, it helps improving test sensitivity by identifying epitopes that trigger strong immune responses, leading to higher detection rates and facilitating timely interventions. Additionally, epitope prediction is cost-effective compared to conventional methods, making broader testing more feasible, especially in resource-constrained settings. Lastly, it accelerates test development during outbreaks of emerging infectious diseases, streamlining the process by focusing on the most promising epitopes [(Jiang et al., 2023); (Campelo et al., 2023)].

Epitope prediction methods also play a crucial role in vaccine development, providing a time- and cost-effective approach to identify peptides capable of enhancing the immune response against infectious agents. Over centuries, vaccines have been instrumental in disease prevention and treatment, leading reduced mortality and morbidity rates, and improved human life expectancy (Rodrigues and Plotkin, 2020). However, traditional vaccine development methods are often time-consuming and costly. In the postgenomic era, the challenge lies in identifying antigenic regions or epitopes that can effectively stimulate the immune system (Parvizpour et al., 2020). Computational *in silico* immunoinformatics provides a solution by aiding rational vaccine design. Epitope-based vaccines have proven to be a promising strategy, delivering both prophylactic and therapeutic effects on pathogen-specific immunity. This approach offers the advantage of eliminating undesirable immune responses, generating prolonged immunity, and remaining cost- and time-effective. The foundation of this strategy lies in the identification of immunodominant epitopes, which trigger immune responses by engaging B cell epitopes (BCEs) or T cell receptors (TCEs) with antigens. By taking advantage of a systematic, computational approach known as *immunoinformatics*, researchers can accelerate vaccine development, leading to more potent disease prevention and treatment methods. This advancement opens up new possibilities for targeted and effective approaches to combat a wide range of diseases (Parvizpour et al., 2020).

From a Computer Science perspective, B-cell epitope prediction is a complex problem due to the diverse nature of epitopes and variability among different organisms. Epitopes are specific segments of proteins that the immune system recognizes and binds to, triggering an immune response. Identifying these regions in proteins is crucial for vaccine development, therapies, and diagnostics. However, accurate prediction of B-cell epitopes remains challenging due to several reasons:

- Epitope diversity: B-cell epitopes can vary in length, amino acid composition, and spatial conformation. Developing algorithms capable of capturing this diversity is

a complex task;

- Scarcity of labeled data: Obtaining accurate labeled data to train epitope prediction models can be difficult and expensive, specially in less-studied organisms or novel pathogens. The traditional supervised learning approach may be limited due to the scarcity of large enough training sets, particularly when trying to develop and/or optimize models for specific pathogens or groups of pathogens.;
- Generalization across taxonomic levels[4]: Transferring knowledge between different taxonomic levels is challenging since proteins, and consequently epitopes, can differ significantly among them.

This thesis proposes a method of improving linear B-cell epitope detection by transfer-learning from higher to lower taxonomic levels using taxon-specific pre-trained models. To achieve this goal effectively, advancements in feature representation are essential. These improvements are essential for enhancing the proposed method's ability to transfer knowledge across taxonomic levels and improve the prediction of linear B-cell epitopes.

1.1 Problem Definition

Identification of linear B-cell epitopes is modeled as a binary classification problem. More precisely, given a protein $\mathcal{A}$ represented as a string of letters representing amino-acids[5] of length r, a model is trained to predict the binary label for each position of the protein, a_i, flagging if the amino-acid is part of a B-cell epitope or not.

Given: A protein $\mathcal{A}$ in the form of a sequence of amino-acids $(a_1, a_2, a_3, ..., a_r)$.

Return: A sequence of labels assigned to each amino-acids $(l_1, l_2, l_3, ..., l_r)$, where l is a binary label with 1 meaning that the corresponding amino-acid is part of a B-cell epitope and 0 meaning that it is not.

1.2 Contributions

The main contribution of this thesis is aimed towards improving linear B-cell epitope prediction by transfer-learning from higher to lower taxonomic level using taxon-specific pre-trained models. This approach involves using a pre-trained protein model based on ESM-1b, trained on higher taxonomic levels, to improve LBCE prediction tasks at lower taxonomic levels. Furthermore, this study evaluates the benefits of creating taxon-specific

[4]Taxonomy is a hierarchical system used by biologists to classify and organize living organisms (Wiley, 2007). Most modern taxonomies tend to reflect evolutionary relationships, also known as *phylogeny*

[5]Amino acids are the building blocks of proteins (Merriam-Webster, 2023a).

pre-trained models. For each organism, we developed a pre-trained model at higher taxonomic level to predict linear epitope regions at the lower taxonomic level. To create this specific pre-trained model, we fine-tuned ESM-1b [6] using higher taxonomic level datasets, which created a model capable of generating useful features to predict epitopes at a lower taxonomic level. By doing so, models can take advantage of underlying biological features at higher levels and transfer this knowledge to improve epitope predictions at lower levels. This method is specially useful in developing bespoke epitope predictors in scenarios where the amount of training data is limited.

More specifically, this thesis aims to make two contributions:

- Transfer learning across taxonomic levels: Present a method that enables the transfer of epitope knowledge across different taxonomic levels. This approach involves investigating how knowledge from epitopes at higher taxonomic levels can be effectively transferred to lower taxonomic levels. This knowledge transfer is particularly valuable in scenarios with scarce labeled data at lower taxonomic levels;
- Feature fusion: Providing more efficient and informative feature representations to describe proteins and epitopes. This involves data preprocessing techniques, combination of features, and the utilization of latent representations. An essential objective in this context is to develop a comprehensive method that integrates evolutionary, physicochemical, and structural features of amino acids, enhancing the overall feature representation.

The proposed method leads to increased performance in comparison to state-of-the-art methods for LBCE prediction in terms of F1 (Fisher, 1936), AUC (Kimber, 1994) and MCC (Matthews, 1975) scores. So far, the results suggest that transfer-learning across taxonomic levels using taxon-specific pre-trained models significantly enhance the classification of linear B-cell epitopes, which can be particularly useful in predicting epitopes in new and less studied pathogens with limited training data availability.

1.3 Thesis Organization

The remainder of this thesis is structured as follows: Chapter 2 provides an overview of the background information relevant to the study. This chapter discusses various aspects such as the theoretical background of epitope prediction methods, the taxonomic levels, and the use of transformers architecture for proteins. Chapter 3 describes the

[6] ESM-1b is a deep contextual language model that has been trained using unsupervised learning techniques on a vast dataset comprising 86 billion amino acids extracted from 250 million protein sequences, which represent a wide range of evolutionary diversity (Rives et al., 2021).

related works on epitope prediction tasks. Chapter 4 presents a detailed description of the proposed method, providing a comprehensive explanation of the approach developed for this research. Chapter 5 presents the results obtained from a series of experiments conducted using the models derived from the proposed method. Chapter 6 discusses the conclusions to the research questions that guided the experiments and proposes future work to be done in the field.

Chapter 2

Background

This chapter serves as an introduction to the field of epitope prediction, exploring its associated challenges and applications. It offers an overview of traditional methods and deep learning techniques employed in epitope prediction, providing insights into the advancements made in this area. Additionally, it highlights the emerging works that contribute to this research.

2.1 B-cell epitope prediction problem

Epitope prediction serves the primary purpose of helping to develop molecules that can serve as substitutes for complete antigens in the production or detection of antibodies. These molecules can be synthesized or generated by cloning the corresponding complete antigen into an expression vector, particularly in the case of proteins. The advantage of using only the epitope, or antigenic determinant, instead of full antigens or weakened or inactivated pathogens (in the case of vaccines) is their cost-effectiveness and non-infectious nature, unlike viruses or bacteria that pose potential risks to researchers, test animals, or individuals with weakened immune systems (Ponomarenko and Regenmortel, 2009).

Synthetic peptides can reproduce both continuous and discontinuous epitopes found on proteins, allowing them to effectively bind to specific antibodies. Continuous epitopes replicate a short sequence, while discontinuous epitopes involve residues from arbitrarily distant protein regions that are brought together in three-dimensional space by protein folding. These peptides can be used for detecting antibodies related to infections, allergies, autoimmune diseases, and cancers (Fleri et al., 2017). Epitope prediction methods are employed to identify peptides capable of binding to specific antibodies. These peptides can also be used to generate antipeptide[1] antibodies for diagnostic purposes. Antibodies play a vital role in detecting proteins and disease markers, especially in early-stage in-

[1] An antipeptide is an antibody that specifically recognizes and binds to a synthetic peptide.

fections. Peptides as short as 10 to 15 amino acids can trigger antibody production, but the challenge lies in finding peptides that have specific antibodies capable of binding to them. Successful prediction of such peptides is crucial for diagnostic design and vaccine development (Ponomarenko and Regenmortel, 2009).

The design of synthetic vaccines is greatly influenced by epitope prediction and identification methods. Some of the current vaccine technologies rely on live attenuated or inactivated pathogens, which require very high levels of quality assurance to ensure that the target organisms are adequately processed in the final product (Zhang and Ulery, 2018). Moreover, attenuated vaccines offer a risk, albeit small, of infection for people with weakened immune systems. In contrast, vaccines based on synthetic peptides employ methods that predict immunogenic peptides capable of eliciting antibodies that neutralize pathogens. This enables researchers to pursue rational vaccine design, which is more cost effective (Piccaluga et al., 2022). However, their understanding of how the immune system specifically responds to various pathogens remains limited. This lack of knowledge makes it challenging to predict which peptides are likely to possess cross-neutralizing immunogenicity, providing effective protection against the target pathogen. Consequently, synthetic vaccine candidates must undergo experimental testing to confirm their ability to generate neutralizing antibodies. Despite the potential benefits, further research and empirical evaluation are necessary to ensure the efficacy and safety of synthetic vaccines (Ponomarenko and Regenmortel, 2009).

In addition, epitope prediction methods are also extensively employed as pre-screening tools to guide experimental investigations. These computational techniques have significantly enhanced the efficiency and cost-effectiveness of B-cell epitope discovery. By utilizing these methods, researchers are able to streamline their experimental efforts and concentrate on a more targeted selection of potential epitopes. This approach not only saves valuable resources but also accelerates the overall discovery process, facilitating the identification of relevant epitopes for further study (Ashford et al., 2021).

B-cell epitope prediction can be categorized into two groups: linear and conformational. Linear epitopes refer to adjacent amino acid sequences in the antigenic sequence, while conformational epitopes consist of amino acids that are separated in the sequence but are brought together through folding. The methods used for predicting B-cell epitopes vary depending on the type of epitope. Although conformational epitopes[2] are more common, most prediction methods focus on linear epitopes due to the scarcity of data on antigen 3D structures and the high computational costs associated with predicting such structures. Linear epitopes can be predicted from amino acid sequence data alone

[2]A conformational epitope is a region of a protein that is recognized by the immune system only when the protein is folded into its correct three-dimensional structure (Shashkova et al., 2022)

and are more stable than conformational epitopes, making them the preferred choice for transportation and storage of potential peptide vaccines. According to (Forsström et al., 2015), linear epitopes show consistent recognition in the sera of rabbits immunized with recombinant proteins and peptides. Additionally, the effects of mutations are easier to estimate for linear epitopes, as most relevant changes occur in the antigenic region. On the other hand, conformational epitopes are more susceptible to alterations caused by amino acid changes in other regions of the protein, resulting in unpredictable, and challenging to model, conformational changes (Pandurangan and Blundell, 2019).

2.1.1 Biological Taxonomies

This thesis aims to address the B-cell epitope prediction problem through transfer-learning, which leverages knowledge from higher to lower taxonomic levels. Therefore, an understanding of biological taxonomy is essential.

In biology, a taxonomy is a hierarchical system used to classify and organize living organisms based on their characteristics and evolutionary relationships (Wiley, 2007). The classification system starts with the hightest level, which is the domain, and goes down to the most specific level, which is the species.

The eight main taxonomic levels are:

- Domain: This is the hightest level of classification, which separates living organisms into three groups based on their cellular structure, biochemical composition, and genetic material. The three domains are Bacteria, Archaea, and Eukarya (Woese et al., 1990);
- Kingdom: Each domain is further divided into Kingdoms based on their cell structure, mode of nutrition, and reproduction. There are currently six kingdoms recognized: Animalia, Plantae, Fungi and Protista (under the domain Eukarya), Archaea, and Bacteria (Wiley, 2007);
- Phylum/Division: Each Kingdom is grouped into Phyla (for animals) or Divisions (for plants and fungi), which groups organisms based on their body plan, cell structure, and other physical characteristics. For instance, the Kingdom Bacteria contains the phylum Pseudomonadota, which includes a large and varied group of bacteria that are found in many different environments (Margulis and Chapman, 2009);
- Class: Each phylum or division is organized into classes, which group organisms based on their physical characteristics, such as the presence of specific organs, body

symmetry, or type of leaves (Cleveland P. Hickman et al., 2017). For example, the Pseudomonadota phylum includes the Betaproteobacteria class;

- Order: Each class is then divided into orders, which group organisms based on their physical characteristics and behavior (Cleveland P. Hickman et al., 2017). For instance, the Betaproteobacteria class contains Burkholderiales order;
- Family: Group organisms based on morphology of their teeth, reproductive organs, or brain size. (Campbell et al., 2021). As an example, the Burkholderiales order includes families such as Alcaligenaceae;
- Genus: Each family is grouped into genera, which divide organisms based on their physical and genetic characteristics (Gupta et al., 2018). For example, the Alcaligenaceae family includes the genera Bordetella;
- Species: Finally, at the species level, organisms are classified into groups that can interbreed and produce viable offspring. Each species is assigned a unique scientific name comprising two parts: the genus and species names (Okasha, 2019). For instance, the scientific name *Bordetella pertussis* (the bacteria that causes whooping cough) has Bordetella as the genus and pertussis as the specific name.

Note that the classification above only applies to cellular organisms. Viruses belong to a completely separate branch and are organized into taxonomic groups arranged in a hierarchical manner, with several primary ranks being used, such as realm, kingdom, phylum, class, order, family, genus, and species. Additionally, there are secondary ranks that exist between the primary ranks, such as subkingdom, subphylum, superfamily etc. (Gorbalenya et al., 2020)

Classifying living organisms based on taxonomic levels has been an important tool for biologists to understand the diversity of life on Earth and their evolutionary relationships. This taxonomic level of an organism can be used to predict the presence and location of B cell epitopes within its proteins. This is because the physical and chemical properties of an antigen are influenced by its evolutionary history, and closely related organisms may have similar antigens with similar epitopes. For example, two species within the same genus (at the taxonomic level of the genus) are likely to have similar proteins and therefore similar B-cell epitopes. Likewise, organisms within the same order or family (at the taxonomic levels of order and family) are likely to share some epitopes. However, as we move up the taxonomic hierarchy towards higher levels, such as superkingdom or phylum, the likelihood of shared epitopes decreases (da Silva et al., 2023).

Therefore, knowing the taxonomic location of a given pathogen on the phylogenetic tree is potentially informative when making predictions about its B-cell epitopes, which

are useful in designing vaccines, diagnostics, and immunotherapeutic formulations based on these epitopes. Moreover, this knowledge contributes to comprehending the interaction between organisms and their immune reactions from an evolutionary perspective. For instance, viruses that are taxonomically similar are likely to share homologous B-cell epitopes, meaning that a vaccine targeting the epitopes of one virus may potentially shield against the other virus as well, as is the case, e.g., of the vaccine used against several viruses in the orthopoxvirus genus (such as smallpox, monkeypox and others), all of which are based on the closely-related *vaccinia* virus (Centers for Disease Control and Prevention, 2023). However, as far as we are aware, very few works to date have taken advantage of these phylogenetic considerations when developing LBCE predictors (Ashford et al., 2021).

2.2 Physicochemical properties-based methods

In the early days of epitope prediction, researchers focused on evaluating individual physiochemical properties of amino acids to identify potential epitopes. They examined properties such as flexibility (Karplus and Schulz, 1985), surface accessibility (Emini et al., 1985), hydrophobicity (Levitt, 1976), and antigenicity (Kolaskar and Tongaonkar, 1990). Researchers developed algorithms that utilize sliding windows along the protein sequence to calculate average amino acid propensity scales. Regions of the protein that scored above a certain cut-off on these scales were identified as potential linear B-cell epitopes. However, it was later determined that relying solely on 484 propensity scales is not reliable enough for accurately detecting BCEs (Blythe and Flower, 2005). To address the limitations of using individual physiochemical properties and propensity scales, more advanced epitope prediction methods have been developed. These methods employ a variety of approaches, including sequence-based algorithms, structural modeling, and machine learning techniques.

2.3 Machine Learning methods

Machine learning (ML) methods have emerged as a powerful tool for predicting linear B-cell epitopes in proteins. These methods rely on multiple propensity scales and incorporate additional amino acid features that were not previously considered (Yang and Yu, 2009). Some examples of popular tools that utilize machine learning methods for B-cell epitope prediction are BepiPred (Larsen et al., 2006a), ABCPred (Saha and Raghava, 2006), LBTope (Singh et al., 2013), APCPred (Shen et al., 2015), iBCE-El (Manavalan et al., 2018), BepiPred 2.0 (Jespersen et al., 2017), DLBEpitope (Liu et al., 2020), Epi-

Dope (Collatz et al., 2020), and EpitopeVec (Bahai et al., 2021). Despite clear progress over the original propensity scale-based methods, these approaches still suffer from inadequate performance when tested across different organism datasets, and they also require a substantial amount of epitope data to train the models. As a notable example, (Ashford et al., 2021) demonstrated promising results in predicting epitopes related to the Epstein-Barr virus (EBV). The method achieved an accuracy (ACC) of 0.72, an area under the receiver operating characteristic curve (AUC) of 0.74, and a Matthews correlation coefficient (MCC) of 0.32.

Deep learning techniques are becoming increasingly common for analyzing proteins, with transfer learning emerging as a notable approach. This method involves using latent space vector representations of amino acid residues that are extracted from large, pre-trained protein language models. These representations have the ability to encode structural, functional, and physicochemical properties of proteins in a context-dependent manner (Chowdhury et al., 2021; Clifford et al., 2022; Elnaggar et al., 2021; Rives et al., 2021), making them a compelling option for developing new models to capture the immunogenic properties of amino acid residues. Transfer learning with protein language models allows researchers to efficiently train models to recognize patterns and features associated with immunogenicity by leveraging pre-existing knowledge of protein structure and function. This approach saves time and computational resources compared to training models from scratch. It enables the development of more accurate prediction models that can identify immunogenic regions in proteins, supporting the design of safer vaccines. BepiPred3.0 stands out as a remarkable deep learning technique in this context. It represents a robust method designed specifically for predicting B-cell epitopes - those specific regions of proteins that the immune system recognizes. By harnessing the power of deep learning algorithms, BepiPred3.0 effectively analyzes the sequence of amino acids in proteins, allowing it to discern potential epitope regions with high accuracy (Clifford et al., 2022).

In summary, the increasing adoption of deep learning techniques, specially transfer-learning, in protein analysis demonstrates the potential of utilizing latent space vector representations obtained from pre-trained protein language models. These representations enable the encoding of various properties of proteins in a context-dependent manner, making them valuable for capturing the immunogenic properties of amino acid residues. This approach holds promise for advancing our understanding of immunogenicity and enhancing the development of protein-based vaccines.

2.4 Transformer model

The hypothesis under consideration is whether language models can effectively address the epitope prediction challenge outlined in the previous chapter. It's worth noting that a significant breakthrough occurred in 2017 with the introduction of the Transformer model by (Vaswani et al., 2017a). This model not only revolutionized the field of language translation but also consolidated its position as the cutting-edge technology in this domain. The model is an encoder-decoder architecture in which the encoder maps the input text's vector representations, referred to as input embeddings[3], to an internal representation. The decoder then employs the internal representation to map it to the output sequences, such as the target language (Figure 2.1). Rather than relying on traditional methods, which employ recurrent or convolution layers, the model uses an attention module[4] to facilitate learning connections between input tokens[5] and interactions between all pairs of tokens in a sequence. The attention module assigns importance weights to each input token for the prediction task, enabling the model to learn dependencies between distant tokens (Cheng et al., 2021). Instead of using a single attention module, which may not be able to capture all the complexities of the relationships between tokens, multiple attention modules are applied in parallel. This approach enables the model to learn multiple different aspects of the relationships between input tokens, allowing for more accurate representations.

The Transformer model includes six layers of both the encoder and decoder blocks. Each block consists of multihead attention with eight parallel attention heads and fully connected feed-forward networks, which act as intermediate components in the overall Transformer model. The input embeddings are generated using two encoding methods: byte-pair encoding (Britz et al., 2017) and word-piece vocabulary (Wu et al., 2016). The model's embedding layers produce contextualized embeddings with a size of 512 per token. By using multihead attention in various layers of each block, the Transformer model acquires valuable representations by taking into account the input sequence's token information from different positions.

Two self-supervision approaches can be utilized for training a Transformer model: predicting the next token in a sequence or predicting masked tokens. The masked language modeling approach has gained popularity for its ability to consider the entire input sequence. The BERT model (Devlin et al., 2019), which uses both self-supervision approaches, has achieved state-of-the-art performance on various NLP tasks, demonstrating

[3]Embeddings refer to the representation of words, phrases, or other entities in a continuous vector space.

[4]It allows models to focus on specific parts of input data, giving more weight to certain elements based on their relevance to the task.

[5]Tokens are the individual units that a text or sentence is divided into.

the importance of bidirectional pretraining[6]. Transformers trained using the masked language modeling (MLM) approach have also gained attention for applications in computational biology and bioinformatics [(Rives et al., 2021); (Elnaggar et al., 2021); (Brandes et al., 2021)].

The Transformer model uses an attention mechanism that allows each input token to affect the weights of every other token in the sequence (Dehghani et al., 2019). This mechanism allows for the incorporation of long-range dependencies within the input sequence, resulting in enhanced sequence embeddings and improved performance (Väth et al., 2022). The Transformer model's direct connections between distant tokens make it more feasible to train and highly parallelizable, resulting in better computational efficiency (Dai et al., 2019).

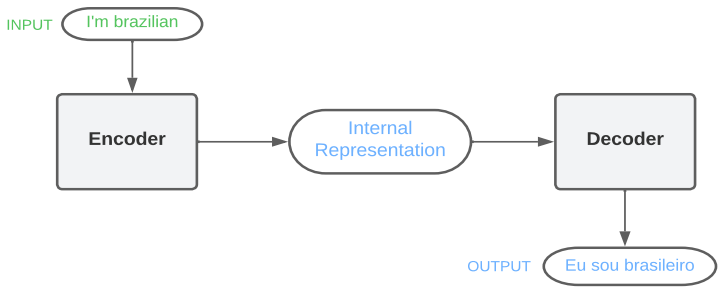

Figure 2.1: An example of how to use the internal representations for downstream machine learning tasks in an encoder-decoder architecture.

2.5 Vanilla Transformer

In this subsection, the key modules of encoder-decoder structure are presented based on Transformer paper (Figure 2.2).

Encoder and Decoder Stacks

As previously mentioned, the encoder consists a series of six identical layers stacked together, each with two sub-layers. The first is a multiple-headed self-attention mechanism, while the second is a position-wise fully connected feed-forward network. A residual connection (He et al., 2016) around each of the two sub-layers is applied, followed by a normalization layer (Ba et al., 2016). The decoder also has a stack of 6 identical layers. The decoder inserts a third sub-layer in addition to the two sub-layers in each encoder

[6]Bidirectional pretraining improves model performance by training language models to predict words in text sequences using both prior and subsequent context.

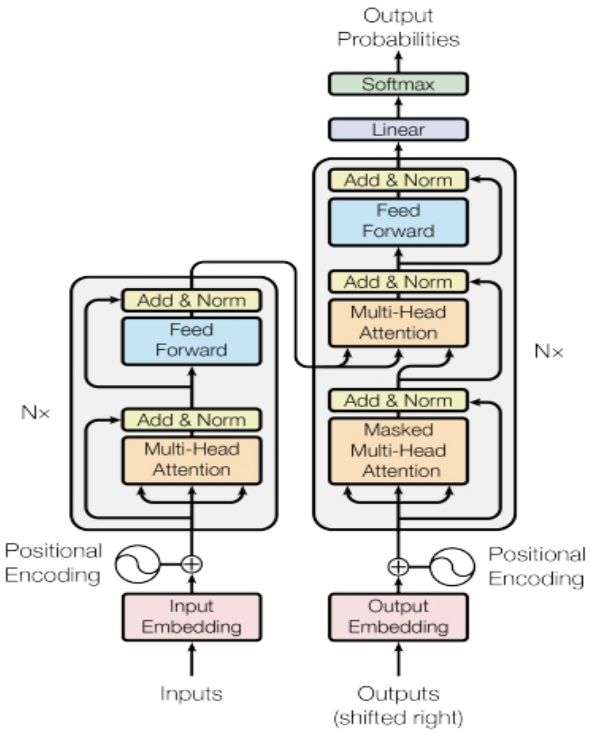

Figure 2.2: The vanilla Transformer architecture (Vaswani et al., 2017b).

layer, which performs multi-head attention over the output of the encoder stack, as shown in the right halves of Figure 2.2 . Here, residual connections are also used around each of the sub-layers, followed by the normalization layer. The self-attention sub-layer is adjusted to avoid subsequent positions.

Self-Attention

Self-attention is a mechanism that enables methods to selectively focus on specific parts of an input sequence, capturing long-range dependencies and relationships between elements (Lin et al., 2017). Compared to recurrent layers, which process input sequences sequentially, self-attention considers the entire sequence when computing the weights for each element, making it more effective at capturing long-range dependencies. Additionally, self-attention can be parallelized across all elements in the input sequence, making it more efficient and scalable than recurrent layers. Furthermore, self-attention is more interpretable than other NLP models because the attention weights computed by the mechanism can be visualized.

In contrast to CNN, self-attention is able to capture relationships between non-adjacent elements in a sequence, whereas CNNs are limited to capturing local relationships between adjacent elements, making them less effective at capturing long-range dependen-

cies (Vaswani et al., 2017a). Therefore, self-attention has become a common choice for handling input sequences of arbitrary length in NLP applications.

Multi-Head Attention

(Vaswani et al., 2017b) propose a method called Multi-Head Attention, which involves linearly projecting the keys[7], values[8], and queries[9] multiple times using different learned linear projections. This is done instead of using a single attention function with keys, values, and queries. The projected copies of the queries, keys, and values are then used in parallel to compute v-dimensional output values. These values are concatenated and projected again to obtain the final output. By using multiple heads, the model is able to attend to data from several sub-spaces of the input sequence simultaneously, allowing it to capture more complex relationships between elements in the sequence.

Feed-Forward Networks

In the encoder and decoder of the Transformer architecture, the feed-forward network (FFN) plays an important role. Comprising of two linear transformations with a ReLU activation function in between, the FFN takes input from the multi-head attention module and applies a non-linear transformation to it. This increases the model's ability to capture complex interactions between input and output sequences (Vaswani et al., 2017a).

Positional Encoding

The Transformer model does not rely on recurrent or convolutional operations, and therefore needs to incorporate information about the positions of tokens in the input sequence to properly understand its order. This is achieved by using of positional encodings, which are added to the input embeddings at the base of both the encoder and decoder stacks. By including these encodings, the model is able to differentiate between the positions of tokens in the input sequence (Vaswani et al., 2017a)

2.6 Protein language models

The utilization of natural language processing techniques in protein language models to comprehend protein sequences is experiencing a notable rise. Although there are similarities between sentences and protein sequences, there are also significant differences in their

[7]Keys represent the elements in the input sequence that one wants to focus on.

[8]Values: Values are the associated information or features corresponding to the elements in the input sequence.

[9]Queries: They are the elements for which relevant information are sought

properties, syntax and semantics. In contrast to language, where words are made up of letters and spaces, proteins consist of individual amino acids or groups of amino acids, with sequences of proteins resembling sentences. The presence of long-range dependencies in protein sequences makes them ideal for analysis using NLP models like Transformers (Heinzinger et al., 2019b); (Ofer and Linial, 2021).

Figure 2.3 shows how a Transformer language model can be used for protein sequences. The encoder maps the amino acid tokens from an input protein sequence to an internal representation known as the protein sequence embedding. The internal representation serves as a feature vector that effectively encapsulates the protein sequence, enabling its utilization as input for conventional machine learning tasks, such as classification or regression (Chandra et al., 2023).

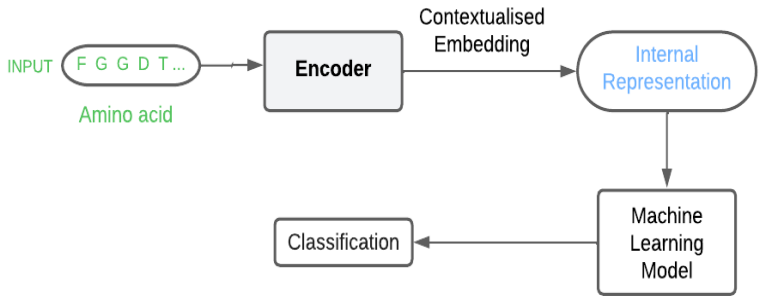

Figure 2.3: This is an example of applying the Transformer language model to predict protein properties. The input is contextualized using the encoder block, which provides an internal representation. This internal representation is then used as amino acid features and is passed to a machine learning model.

For many years, researchers have been working on predicting the structure of proteins, which is determined by their amino acid composition and plays a crucial role in determining their function (Jumper et al., 2021a). This task is divided into two categories: predicting the secondary structure (α- helix, β-sheet, or coil) and predicting the tertiary structure (3D shape). To achieve this, other tasks such as predicting 2D contacts can be performed, which can be used to further predict the 3D structure.

Recent studies have shown promising results in predicting secondary structure and contact using different Transformer models. For instance, (Rives et al., 2021) achieved high performance in predicting secondary structure by training a neural network classifier using sequence profile capabilities combined with the ESM-1b Transformer model. They evaluated the performance of their model using the Critical Assessment of Protein Struc-

ture Prediction (CASP)[10] test set (Kryshtafovych et al., 2019) and demonstrated better results compared to other models. Another example worth mentioning is AlphaFold 2, a cutting-edge deep learning model specifically designed to achieve precise predictions of the 3D structure of proteins (Jumper et al., 2021b).

2.7 Conclusion

This chapter discusses various aspects of B-cell epitope prediction, which is important for the development of molecules that can be used in antibody production and vaccine design. It explains that synthetic peptides can mimic epitopes and be used for developing vaccines. The text also mentions the use of epitope prediction methods as pre-screening tools and the categorization of epitopes into linear and conformational types. The chapter also describes the use of physicochemical properties-based methods and machine learning methods, particularly deep learning and transfer-learning for epitope prediction. Additionally, the text discusses the concept of taxonomic levels in biology and how they can be used to predict the presence and location of B-cell epitopes in proteins. It explains that closely related organisms are likely to have similar epitopes, while the likelihood decreases as we move up the taxonomic hierarchy.

Finally, the chapter introduces the Transformer model, an encoder-decoder architecture that has shown remarkable progress in language translation. It explains how the model uses attention mechanisms to capture dependencies between input tokens and achieve more accurate representations. The chapter also emphasizes the training approaches for Transformer models, such as predicting the next token or masked language modeling, and the use in protein language models.

[10]The Critical Assessment of protein Structure Prediction (CASP) is a community-wide experiment held periodically to evaluate the state-of-the-art methods in predicting protein structures (Moult et al., 2017).

Chapter 3

Related Work

This chapter provides a literature review to comprehensively summarize recent research and developments in epitope prediction, deep learning, and related fields. The review also includes a critical assessment of research findings to highlight their strengths and limitations.

3.1 Introduction

The concept that the statistical patterns present in protein sequences hold crucial information about their biological function and structure is grounded in scientific research, as demonstrated by previous studies [(Yanofsky et al., 1964); (Altschuh et al., 1987)]. During evolution, sequences that correlate with the most favorable fitness[1] outcomes are typically selected among a vast range of random perturbations that are possible (Göbel et al., 1994). These unobservable factors that dictate a protein's contribution to fitness, such as its stability, structure, and function, are reflected in the distribution of naturally occurring sequences that can be observed (Göbel et al., 1994).

The decoding of the information contained in protein sequence variation is a long-standing issue in biology. In the field of machine learning, natural language understanding is similar, with the distributional hypothesis suggesting that word semantics can be inferred from their contextual usage (Harris, 1954).

Recently, self-supervised learning has emerged as a main direction in machine learning research. Self-supervised methods use unlabeled datasets, such as predicting the next word in a sentence or masked words in a context, rather than manual annotation, allowing them to exploit significantly larger amounts of data [Bengio et al. (2003); Devlin et al. (2019)].

[1]The biological term "fitness" describes an organism's ability to survive and reproduce in its environment based on its genetic features. It serves as a measure of the organism's genetic contribution to the next generation (StudySmarter, 2023).

Recent studies have shown that self-supervised methods, combined with large data and high-capacity models, have produced new state-of-the-art results, such as approaching human performance in question answering and semantic reasoning benchmarks, and deep learning for protein modelling [Devlin et al. (2019); Rives et al. (2021)].

The literature review methodology followed the best practices outlined in"Guidelines for performing systematic literature reviews" by (Kitchenham and Charters, 2007). This involved using keywords such as "epitope prediction," "transfer-learning," "protein modeling," and "self-supervised learning" to search the CAPES Portal of Scientific Publications ("Portal CAPES"), which aggregates metadata from all the major publishers including Springer, Elsevier, ACM, OUP, IEEE press and others. The search was initially filtered to knowledge areas related to Computer Science. Then, the selection was refined to focus on papers from sources with a high CAPES Qualis rating (A1 or A2) and a substantial number of citations. However, some recent papers with fewer citations were selected ad-hoc due to a strong alignment with the research scope.

3.2 Deep Learning for Protein Modelling

(Rives et al., 2021) investigate the application of the Transformer neural network architecture for modeling large datasets of amino acid sequences. They demonstrate the effectiveness of their approach by utilizing UniRef[2] to construct three pre-training datasets with varying levels of sequence diversity. The datasets used contain 250 million protein sequences totalling 86 billion amino acids, which are comparable in size to the large text datasets commonly used in natural language processing [Devlin et al. (2019); Radford et al. (2019)]. The authors utilize the Transformer architecture in their experiment due to its remarkable performance in natural language processing tasks. They implemented a deep Transformer that takes amino acid character sequences as input and processes them through a sequence of blocks that alternate self-attention with feed-forward connections. For training the models, the authors employed a masked language modeling objective. The study concludes with the development of a novel pre-trained model known as ESM-1b, which is trained on a high sequence diversity dataset with approximately 650 million parameters. The ESM-1b surpasses all previously tested models, implying that further improvements in performance could be achieved by using even higher-capacity models. Overall, the study demonstrates that the sequence diversity in pre-training data plays a crucial role in shaping the performance of pre-trained models for protein language modeling.

[2]UniRef (UniProt Reference Clusters) is a database that clusters protein sequences into groups based on their sequence similarity.

In the wider context of applying Deep Learning techniques for protein analysis, one topic that has received a great deal of attention in recent years is 3D structure prediction, since the publication of AlphaFold (Jumper et al., 2021b). Its updated version, AlphaFold2, stands as a groundbreaking protein structure prediction method, surpassing its competitors with remarkable precision and accuracy. The system's success is attributed to novel neural network architectures and training procedures based on evolutionary, physical, and geometric constraints of protein structures. AlphaFold2 utilizes amino acid sequences to create a Multiple Sequence Alignment (MSA) and identifies mutation-prone regions and correlations between them. It also identifies proteins with similar structures to build an initial representation (template) called pair representation. The key components of AlphaFold2 are the evoformer and structure module, both based on attention mechanisms. The evoformer exchanges information between MSA and templates to improve assessment and align them correctly. It consists of two specialized transformers for MSA and pair representations. The structure module uses both representations to prioritize protein backbone orientation, considering residue rotations and translations, and performs local refinement and minimization using gradient descent (Bertoline et al., 2023). This breakthrough in neural network architectures makes AlphaFold2 highly effective in predicting protein structures.

(Lin et al., 2022) introduce a method called ESMFold, which enables accurate and end-to-end atomic level structure prediction directly from a protein's primary sequence. The ESMFold model architecture consists of three main components: the ESM-2 language model (Lin et al., 2023), the folding trunk, and the structure module. The language model provides information to the folding trunk, which processes the data, and then passes it to the structure module. This final module is responsible for generating 3D coordinates and confidence values as the output of the model. ESMFold showcases performance levels comparable to other state-of-the-art methods like AlphaFold2 and RoseTTAFold (Baek et al., 2021). In comparison to AlphaFold2, ESMFold's prediction process is an order of magnitude faster. While AlphaFold2 and and RoseTTAFold have shown breakthrough success in predicting protein structures, they heavily rely on multiple sequence alignments (MSAs) and templates of similar protein structures for optimal performance. In contrast, ESMFold takes advantage of the internal representations learned by the language model and can generate structure predictions using only a single protein sequence as input. This characteristic makes ESMFold significantly faster in predicting protein structures compared to the other models. So far, the present work has not touched 3D protein structure estimation, but it remains as a topic of interest for next steps.

3.3 Deep Learning for Epitope Prediction

Out of all the works analised in our literature search, the most closely related is BepiPred 3.0 (Clifford et al., 2022). The method improves sequence-based epitope prediction tool that uses ESM-1b model from (Rives et al., 2021) to enhance prediction accuracy for linear and conformational epitopes. To this end, they demonstrate that the tool's performance is improved by including extra input variables and refining the epitope residue annotation strategy. The dataset was created by identifying epitope residues in crystal structures containing at least one antibody and one antigen protein chain, and then reducing redundancy through an epitope collapse strategy. The dataset was clustered at different sequence identity thresholds to obtain the final datasets. Additionally, three independent test sets were constructed with updated and enriched epitope annotations, additional antigens, and linear B cell epitopes from the IEDB. The article considers three different methods to represent residues: sparse encoding, BLOSUM62 log-odds scores, and numeric embeddings from the ESM-1b protein language model. They trained three types of neural networks - Feed Forward (FFNN), Convolutional (CNN), and Long Short-term Memory (LSTM) - on sparse, BLOSUM62, and ESM-1b encodings, with or without additional variables. A Random Forest Classifier (RFC) was also trained as a baseline. Residues were represented by concatenating encodings from the residue and its neighboring residues, and the encodings also included a feature corresponding to predicted surface accessibility of different parts of the protein, calculated using NetSurfP 3.0[3] predicted relative surface accessibility (RSA) values for the central residue and protein sequence lengths. Target values were encoded in a position-wise binary manner to distinguish epitope and non-epitope residues.

Another closely related method is EpiDope of (Collatz et al., 2020). The authors develop a python tool that uses a deep neural network to detect linear B-cell epitope regions on protein sequences. They used the IEDB Linear Epitope Dataset to train their deep neural network (DNN) for detecting epitopes. The dataset contains 30,556 protein sequences, each marked with a verified epitope or non-epitope region. The authors preprocessed the dataset to ensure the best possible training basis for their DNN. They first merged identical protein sequences, resulting in a reduced dataset containing 3158 proteins preserving all verified regions. Then they clustered highly similar protein sequences using CD-HIT[4] and an identity threshold of 0.8, and retained only the protein sequence

[3]NetSurfP is a tool for predicting the accessibility of protein residues, with NetSurfP 3.0 using a neural network approach to predict the relative surface accessibility (RSA) of amino acid residues in proteins (Høie et al., 2022).

[4]CD-HIT is a widely recognized and extensively employed program within the field of bioinformatics for clustering biological sequences. Its primary purpose is to reduce sequence redundancy, enhancing the efficiency and accuracy of subsequent sequence analyses (Fu et al., 2012)

with the largest number of verified regions from each cluster, resulting in 24,610 verified regions. In terms of model, the authors tested different DNN architectures to predict epitopes from protein sequences, and established a two-part architecture for their EpiDope model. The first part uses an ELMo DNN to produce context-sensitive embeddings of amino acids, which are then processed by a bidirectional LSTM layer and a dense layer. The second part encodes each amino acid into a non-context-sensitive vector and processes them through another bidirectional LSTM layer and dense layer before being combined with the output of the first part and fed into a final dense layer representing the two classes, epitope and non-epitope. The authors did not fine-tune the ELMo DNN due to the high number of parameters utilized by ELMo, the limited number of samples available for their classification task, and the need to avoid overfitting. As a result, the study claims to achieve a bias-free prediction of epitopes by analyzing a diverse set of known epitopes from evolutionarily distinct organisms in the training set.

Continuing in the field of deep learning applied to epitope prediction, (Bahai et al., 2021) presents a LBCE prediction method called EpitopeVec that utilizes a combination of residue properties, modified antigenicity scales, and protein language model-based representations. They obtained a dataset of viral peptides reported as epitopes and non-epitopes from the IEDB, where peptide length varied from 6 to 46 amino acids. CD-HIT was used to remove homologous sequences, and common peptides between the epitope and non-epitope sets were removed to obtain a final dataset of 4432 positive epitopes and 8460 negative non-epitopes. The authors proposed using distributed vector representations of biological sequence segments, called bio-vectors and ProtVec for proteins, as an alternative to k-mers. They trained a skip-gram neural network on large protein sequence databases to predict surrounding words for a given word (k-mer) and used negative sampling during training to avoid computational expense. After training, they used summation embedding of the existing k-mers in a given protein sequence to represent the sequence, which has proven helpful in protein function annotation tasks. In this study, machine learning (ML) algorithms were employed to differentiate peptides into two categories: epitopes and non-epitopes. For binary classification, Support Vector Machines (SVMs) with the RBF kernel were utilized, providing an approach to distinguish between these two classes. The authors aimed to identify the best performing features for their ML model and trained the classifier using small and large datasets derived from the Bcipep (Saha et al., 2005) and IEDB (Vita et al., 2009). They conducted a performance comparison of their method against state-of-the-art techniques such as those proposed by (Larsen et al., 2006b), (Jespersen et al., 2017), and (Collatz et al., 2020). Notably, their approach yielded superior results in this evaluation.

3.4 Transfer learning in protein data

(Bugnon et al., 2023) state that the automatic annotation of the protein universe remains an unsolved challenge, with only a tiny fraction (0.25%) of the 229,149,489 entries in the UniProtKB database being functionally annotated. Currently, the manual annotation process relies on the protein families database (Pfam), which uses sequence alignments and Hidden Markov Models to annotate family domains. However, this approach has seen slow growth in Pfam annotations over the years. Recently, deep learning models have emerged, capable of learning evolutionary patterns from unaligned protein sequences. Despite their potential, these models face challenges with limited data, as many families have only a few sequences. To address this limitation, they propose utilizing transfer learning, leveraging self-supervised learning on large unannotated data, and then fine-tuning with supervised learning on a small labeled dataset. Their results demonstrate a promising reduction of 55% in errors compared to standard methods for protein family prediction.

(Heinzinger et al., 2019a) introduce a methodology for representing protein sequences as continuous vectors (SeqVec) using the language model ELMo from natural language processing. SeqVec captures the biophysical properties of protein language from unlabeled data, outperforming traditional one-hot encoding and Word2vec-like approaches in predicting secondary structure and regions with intrinsic disorder. They also demonstrate good performance in predicting sub-cellular localization and distinguishing membrane-bound from water-soluble proteins. While SeqVec proves to be the fastest method, it does not surpass the best existing method using evolutionary information. However, SeqVec's speed makes it highly scalable for big data analysis in proteomics, such as microbiome or metaproteome studies. Overall, transfer learning successfully extracts relevant information from unlabeled sequence databases, making SeqVec a powerful tool for various protein prediction tasks, except for cases where evolutionary information is available.

(Shashkova et al., 2022) employ a transfer learning approach using pretrained deep learning models to build a new predictive model, SEMA. By utilizing the ESM-1v (Meier et al., 2021) protein language model and the ESM-IF1 (Hsu et al., 2022) inverse folding model, they fine-tuned them to quantitatively predict antibody-antigen interactions and distinguish epitope and non-epitope residues. SEMA outperformed existing peer-reviewed tools with a reported ROC AUC of 0.76 on an independent test set. Moreover, they demonstrate that SEMA can effectively rank immunodominant regions within the SARS-CoV-2 RBD domain, holding significant potential for advancements in vaccine research and immunotherapy drug development.

3.5 Discussion

BepiPred 3.0 (Clifford et al., 2022) already incorporates the ESM-1b model, leveraging transfer learning to reduce the required amount of specific training data. This predictor is based on datasets containing labeled peptide sequences from various organisms, aiming to develop a general predictor that can be used without specifying the source organism of the submitted peptides. However, as demonstrated by (Ashford et al., 2021), generalist approaches may lead to lower performance. In contrast to BepiPred 3.0, the methodology developed in this thesis differs in a number of key aspects: (1) Transfer learning is employed from higher to lower taxonomic levels, to generate (2) Organism- or taxon-specific pre-trained models instead of general pre-trained models and (3) develop a method that integrates evolutionary, physicochemical, and structural features of amino acids. This work will also (4) experimentally determine the limits of this taxon-specific training approach in terms of minimum data requirements and required computational capabilities.

Chapter 4

Proposed method

We present a method to enhance the generation of features for predicting linear B-cell epitopes from sequence data, by transfer-learning from higher to lower taxonomic levels using taxon-specific pre-trained models. The motivating hypothesis is that knowledge learned from higher taxonomic levels contains valuable information that can improve the prediction performance at lower taxonomic levels. This study incorporates the benefits of creating taxon-specific pre-trained models, which were established in the earlier work of (Ashford et al., 2021) and (Campelo et al., 2023), and advances that work by investigating the ability of transfer-learning strategies to leverage information from taxonomic-related pathogens to improve the performance of bespoke models tailored to specific pathogens.

4.1 Datasets

In this thesis, we made use of three data sources to compose our datasets: Immune Epitope Database - IEDB (Vita et al., 2018), National Center for Biotechnology Information - NCBI (Sayers et al., 2020), and Universal Protein Resource - UniProt (Consortium, 2022)

IEDB is a curated database that focuses on immunology-related data, particularly epitope information related to immune responses. IEDB was established in 2004 and contains over 1.6 million experiments related to the adaptive immune response to epitopes. The data is manually curated (Vita et al., 2008) from the scientific literature, primarily from PubMed (Vita et al., 2014). The IEDB team has curated all the literature available from the beginnings of PubMed until 2011, and they continue to update the database with newly published papers every two weeks.

NCBI dataset is a collection of biological data and literature that is freely available to the public. It is one of the most comprehensive biological databases in the world and is used by researchers and scientists to access and analyze biological data.

UniProt is a widely used protein database that provides a vast repository of information on proteins from various organisms. The goal is to provide a standardized and centralized source of protein sequence and functional information, helping researchers, scientists and bioinformatics in understanding the roles and characteristics of proteins.

4.2 Data extraction and preparation

The data extraction, filtering, and consolidation process involved using functions from the R package "epitopes" (Campelo and Ashford, 2022). The process was based on the complete XML export of the Immune Epitopes Database (Vita et al., 2019). Specifically, entries classified as Linear B Cell Epitopes (LBCEs) from organisms within the superkingdoms Viruses (NCBI:txid10239), Bacteria (NCBI:txid2) and Eukaryota (NCBI:txi2759) were extracted from the IEDB export. The associated proteins were retrieved from the NCBI protein database (NCBI, 2015) and UniprotKB (UniProt, 2020) [Figure 4.1]. To label each peptide, a positive classification was assigned if at least half of the assays[1] associated with a specific IEDB entry reported a positive result. Positive peptides longer than 30 amino acid residues, which could potentially introduce excessive noise to the training data as they represent lengthy “Epitope-containing regions”, were excluded. Additionally, to avoid duplication of partial information, overlapping peptides belonging to the same class were merged into a single entry. To examine the similarities among the candidates, a range of tools, including BLASTp (Altschul et al., 1990) could be utilized. In this work, the degree of similarity was assessed using the Smith-Waterman similarity (Smith and Waterman, 1981) [Figure 4.1], since the cost of computing the optimal Smith-Waterman alignments for the number of proteins involved in the work is feasible.

The aforementioned process was applied to generate datasets for twelve distinct pathogens (or, in some cases, pathogen-containing lower taxa such as genus or family), which were chosen as case studies so as to provide good diversity of examples across bacterial, viral and eukaryotic pathogens.

[1] An "assay" refers to a set of laboratory techniques and procedures used to measure, analyze, or evaluate a specific biological, chemical, or physical property of a substance

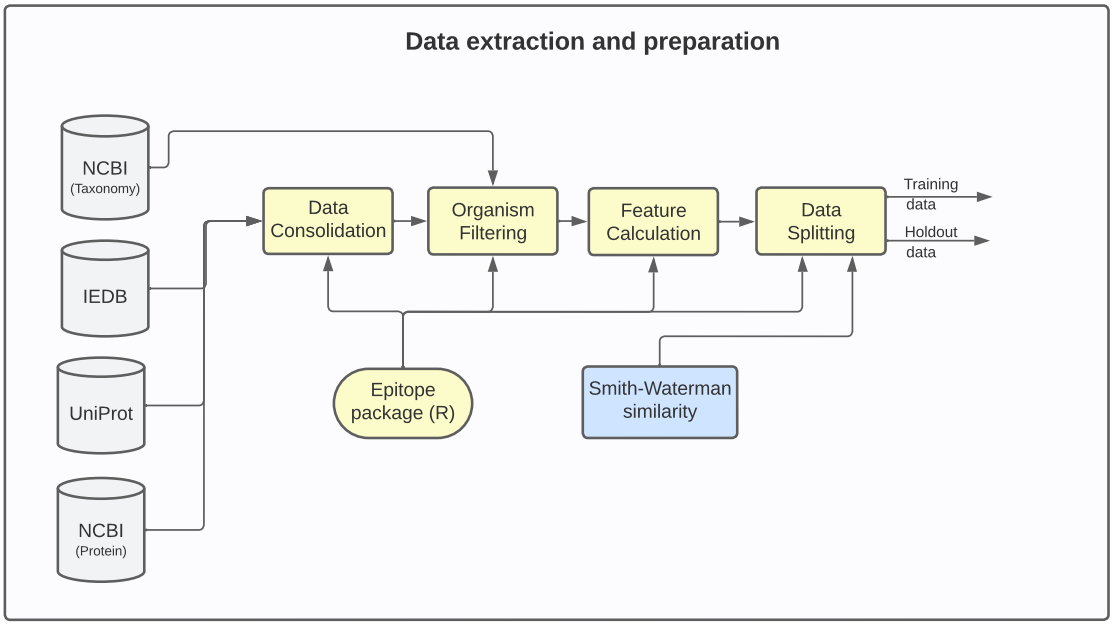

Figure 4.1: Publicly available data is collected from IEDB (Vita et al., 2019), NCBI (NCBI, 2015), and UniProt (UniProt, 2020) to construct an organism-specific dataset. The dataset is subsequently divided at the protein level, using protein ID and similarity, into a training set (utilized for model development) and a hold-out set (utilized to assess the generalization performance of the models). The epitopes R package, which incorporates the key elements of this pipeline, can be found at https://fcampelo.github.io/epitopes.

4.3 Modelling

The model architecture consists of three main components:

1. Fine-tuning (Figure 4.2) involves using ESM-1b as the base model and training it with a dataset from higher-level organisms (Table 4.1). The training data is processed at the amino acid level, utilizing a sliding window of length 1024. This window is moved across each amino acid to extract training samples. By incorporating epitope data from higher-level organism datasets during fine-tuning, the base model can acquire knowledge on both epitopes and non-epitopes.

2. In feature generation (Figure 4.2), the pre-trained model generated at a higher taxonomic level is leveraged to extract features from protein sequences at lower taxonomic level datasets (Table 4.2). These features are used to train a classifier to predict the linear B-cell epitopes. By utilizing the pre-trained model to generate features, the model can integrate the knowledge acquired at higher taxonomic levels

Taxonomic level	**Labelled peptides (-/+)**	**Number of proteins**	**Detail**
Pseudomonadota NCBI:txid1224 (Phylum)	242- / 490+	310	Excluding all *B. pertussis* entries
Terrabacteria NCBI:txid1783272 (Clade)	965- / 875+	619	Excluding all *Corynebacterium* entries
Bamfordvirae NCBI:txid2732005(Kingdom)	25- / 104+	55	Excluding all *Orthopoxvirus* entries
Pseudomonadota NCBI:txid1224 (Phylum)	254- / 457+	265	Excluding all *E. coli* entries
Pseudomonadota NCBI:txid1224 (Phylum)	230- / 398+	234	Excluding all *Enterobacteriacea* entries
Pararnavirae NCBI:txid2732397(Kingdom)	176- / 311+	176	Excluding all *Lentivirus entries*
Terrabacteria NCBI:txid1783272 (Clade)	710- / 566+	419	Excluding all *M. tuberculosis* entries
Pseudomonadota NCBI:txid1224 (Phylum)	264- / 539+	316	Excluding all *P. aeruginosa* entries
Orthornavirae NCBI:txid2732396(Kingdom)	480- / 480+	688	Excluding all *SARS-CoV-2* entries
Platyhelminthes NCBI:txid6157 (Phylum)	147- / 132+	95	Excluding all *S. mansoni* entries
Apicomplexa NCBI:txid1184 (Phylum)	285- / 285+	312	Excluding all *T. gondii* entries
Sar NCBI:txid2698737 (Clade)	151- / 265+	183	Excluding all *P. falciparum* entries

Table 4.1: Datasets used to create pre-trained model at a higher taxonomic level.

Taxonomic level	**Labelled peptides (-/+)**	**Number of proteins**	**Detail**
B. pertussis NCBI:txid520	34- / 61+	16	Split in five folds. Smith-Waterman similarity threshold: 0.7
Corynebacterium NCBI:txid1716	12- / 13+	5	Split in five folds. Smith-Waterman similarity threshold: 0.7
Orthopoxvirus NCBI:txid10242	14- / 20+	15	Split in five folds. Smith-Waterman similarity threshold: 0.7
E. coli NCBI:txid562	22- / 94+	61	Split in five folds. Smith-Waterman similarity threshold: 0.7
Enterobacteriacea NCBI:txid543	46- / 153+	92	Split in five folds. Smith-Waterman similarity threshold: 0.7
Lentivirus NCBI:txid11646	12- / 99+	87	Split in five folds. Smith-Waterman similarity threshold: 0.7
M. tuberculosis NCBI:txid1773	267- / 322+	205	Split in five folds. Smith-Waterman similarity threshold: 0.7
P. aeruginosa NCBI:txid287	12- / 12+	12	Split in five folds. Smith-Waterman similarity threshold: 0.7
SARS-CoV-2 NCBI:txid694009	795- / 274+	195	Split in five folds. Smith-Waterman similarity threshold: 0.7
S. mansoni NCBI:txid6183	243- / 173+	100	Split in five folds. Smith-Waterman similarity threshold: 0.7
T. gondii NCBI:txid5811	60- / 82+	77	Split in five folds. Smith-Waterman similarity threshold: 0.7
P. falciparum NCBI:txid5833	120- / 120+	166	Split in five folds. Smith-Waterman similarity threshold: 0.7

Table 4.2: Datasets used for development and validation of the LBCE predictor of specific organisms at a lower taxonomic level.

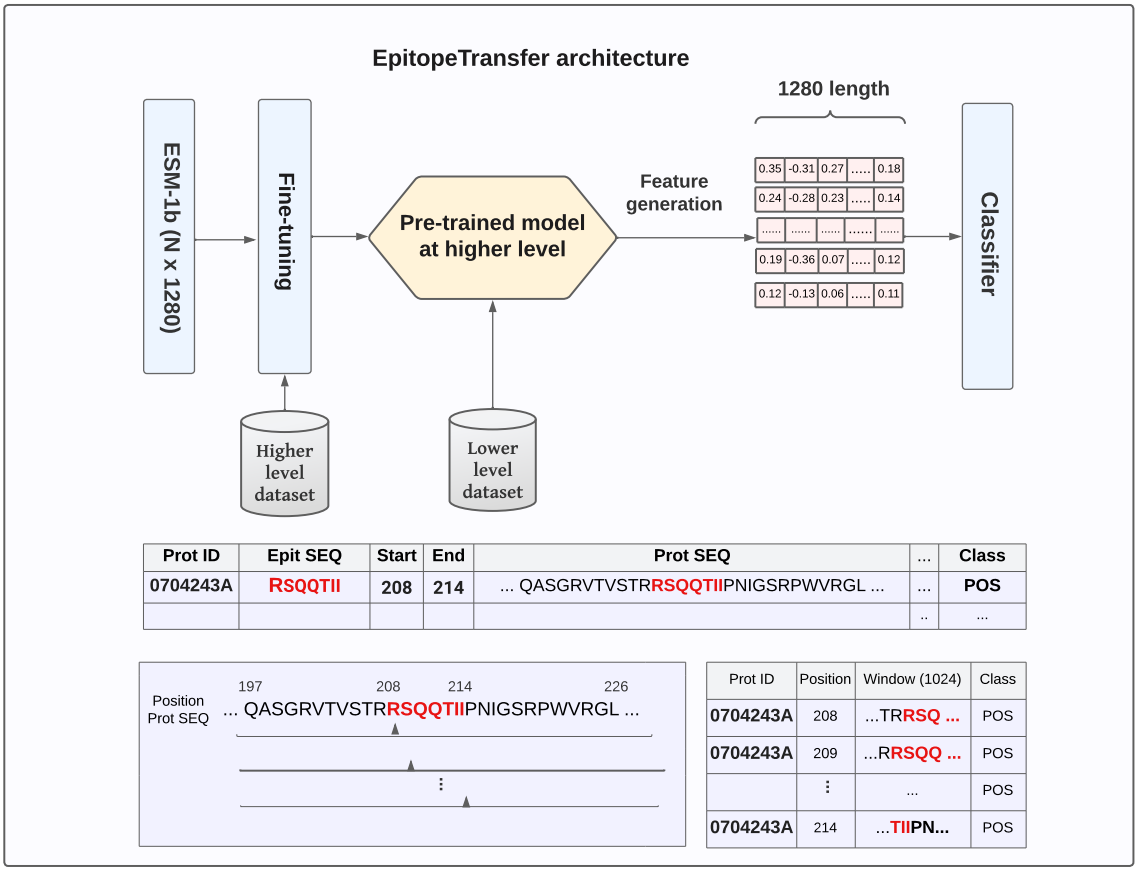

Figure 4.2: *Top*: Overview of the method to create pre-trained protein models based on a higher taxonomic levels, and generate features at lower taxonomic levels for classification tasks. *Bottom*: To generate the training samples at higher level, a 1024-AA sliding window representation with a step size of one is employed. Subsequently, for each amino acid (AA), 1280 features are extracted and labeled accordingly.

and improve prediction performance at lower taxonomic levels. During the feature generation process, the entire protein sequence is inputted into the pre-trained model that is based on a higher taxonomic level. From this output, only the labeled peptide regions are selected for training. This enables the model to capture a richer contextual, resulting in a enhanced feature representation for each amino acid.

3. In the classification step (Figure 4.2), enhanced features generated from the previous phase are utilized to feed any chosen classification method. In this work, the Random Forest (Breiman, 2001) algorithm is used.

4.4 Conclusion

This chapter presented a method to enhance the development of predictive models for identifying linear B-cell epitopes within sequence data. The method leverages transfer-learning, where knowledge gained from higher taxonomic levels is applied to improve predictions at lower taxonomic levels. This improvement is obtained by employing taxon-specific pre-trained models.

Chapter 5

Preliminary Results

The approach described in the previous section was evaluated on the twelve distinct organism datasets previously documented. The proposed method, which we call EpitopeTransfer for the remainder of this chapter, was found to outperform all baseline methods across AUC, F1, and MCC, strongly suggesting that some useful information for the prediction of linear B-cell epitopes is indeed being tranfered down to the lower taxonomic levels (Table 5.1 and Table 5.2). To explore a tentative explanation for the data reasons for this observed increased performance of organism-specific models, we used t-SNE (Van der Maaten and Hinton, 2008) to investigate whether data from different pathogens exhibit different clustering or neighborhood structures in terms of positive/negative observations (Section 5.1). In conclusion, an ablation study was conducted to assess the performance of EpitopeTransfer in comparison to the base model, ESM-1b.

Evaluation data				
Method	Taxon	AUC	F1	MCC
EpitopeTransfer				
	B. pertussis	0.582	0.542	0.263
	Corynebacterium	0.612	0.600	0.253
	Orthopox	0.613	0.591	0.274
	E. coli	0.924	0.857	0.724
	Enterobacteriacea	0.797	0.767	0.559
	Lentivirus	0.793	0.870	0.770
	M. tuberculosis	0.608	0.576	0.196
	P. aeruginosa	0.655	0.595	0.390
	SARS-Cov-2	0.576	0.547	0.163
	S. mansoni	0.531	0.540	0.116
	T. gondii	0.694	0.636	0.308
	P. falciparum	0.759	0.705	0.465
BepiPred 3.0				
	B. pertussis	0.516	0.562	0.279
	Corynebacterium	0.615	0.459	0.140
	Orthopox	0.706	0.646	0.301
	E. coli	0.877	0.690	0.471
	Enterobacteriacea	0.787	0.367	0.077
	Lentivirus	0.467	0.444	-0.111
	M. tuberculosis	0.643	0.324	0.000
	P. aeruginosa	0.225	0.322	-0.355
	SARS-Cov-2	0.451	0.454	-0.025
	S. mansoni	0.609	0.424	0.000
	T. gondii	0.447	0.406	-0.039
	P. falciparum	0.746	0.445	0.077

Table 5.1: A comparison of the F1, AUC, and MCC results between EpitopeTransfer and BepiPred 3.0 is presented. EpitopeTransfer outperforms BepiPred 3.0 in nearly all cases, with the exception of Orthopox. For *M. tuberculosis* and *S. mansoni*, BepiPred 3.0 exhibits superiority only in terms of AUC, while demonstrating inferior results in terms of F1 and MCC. BepiPred 3.0 performance will likely decrease because we have not yet removed samples from our test dataset that were originally used to train the BepiPred 3.0 model.

Evaluation data				
Method	Taxon	AUC	F1	MCC
EpiDope				
	B. pertussis	0.532	0.219	0.000
	Corynebacterium	0.587	0.515	0.284
	Orthopox	0.495	0.499	0.024
	E. coli	0.762	0.263	0.000
	Enterobacteriacea	0.538	0.343	-0.151
	Lentivirus	0.295	0.133	-0.317
	M. tuberculosis	0.627	0.343	0.090
	P. aeruginosa	0.565	0.224	0.000
	SARS-Cov-2	0.425	0.405	-0.137
	S. mansoni	0.561	0.484	-0.013
	T. gondii	0.491	0.236	0.000
	P. falciparum	0.454	0.575	0.214
EpitopeVec				
	B. pertussis	0.546	0.572	0.354
	Corynebacterium	0.745	0.321	0.000
	Orthopox	0.651	0.236	0.057
	E. coli	0.478	0.411	-0.041
	Enterobacteriacea	0.616	0.393	-0.029
	Lentivirus	0.860	0.634	0.422
	M. tuberculosis	0.442	0.369	-0.147
	P. aeruginosa	0.534	0.499	0.193
	SARS-Cov-2	0.746	0.406	0.086
	S. mansoni	0.428	0.326	-0.132
	T. gondii	0.644	0.475	0.121
	P. falciparum	0.564	0.257	0.041

Table 5.2: EpitopeTransfer outperforms both EpiDope and EpitopeVec in most cases. However, it is outperformed by EpiDope in terms of MCC for *Corynebacterium*, AUC for *M. tuberculosis* and *S. mansoni*. It is also outperformed by EpitopeVec in terms of F1 and MCC for *B. pertussis*, and AUC for *Corynebacterium, Orthopox, Lentivirus, and SARS-CoV-2*. The performance of EpiDope and EpitopeVec is expected to decrease since we have not yet excluded samples from our test dataset that were employed to train both models.

5.1 Estimated density

To qualitatively investigate the neighborhood structure of the datasets, we used a t-SNE projection (Van der Maaten and Hinton, 2008) of the whole data, which we later stratified by pathogen group. We aimed to investigate whether positive/negative data from distinct pathogens clustered around distinct regions of the feature space. Insights gathered from this projection could help explain the enhanced performance of taxon-specific models over generalist approaches, without however addressing the underlying biological mechanisms.

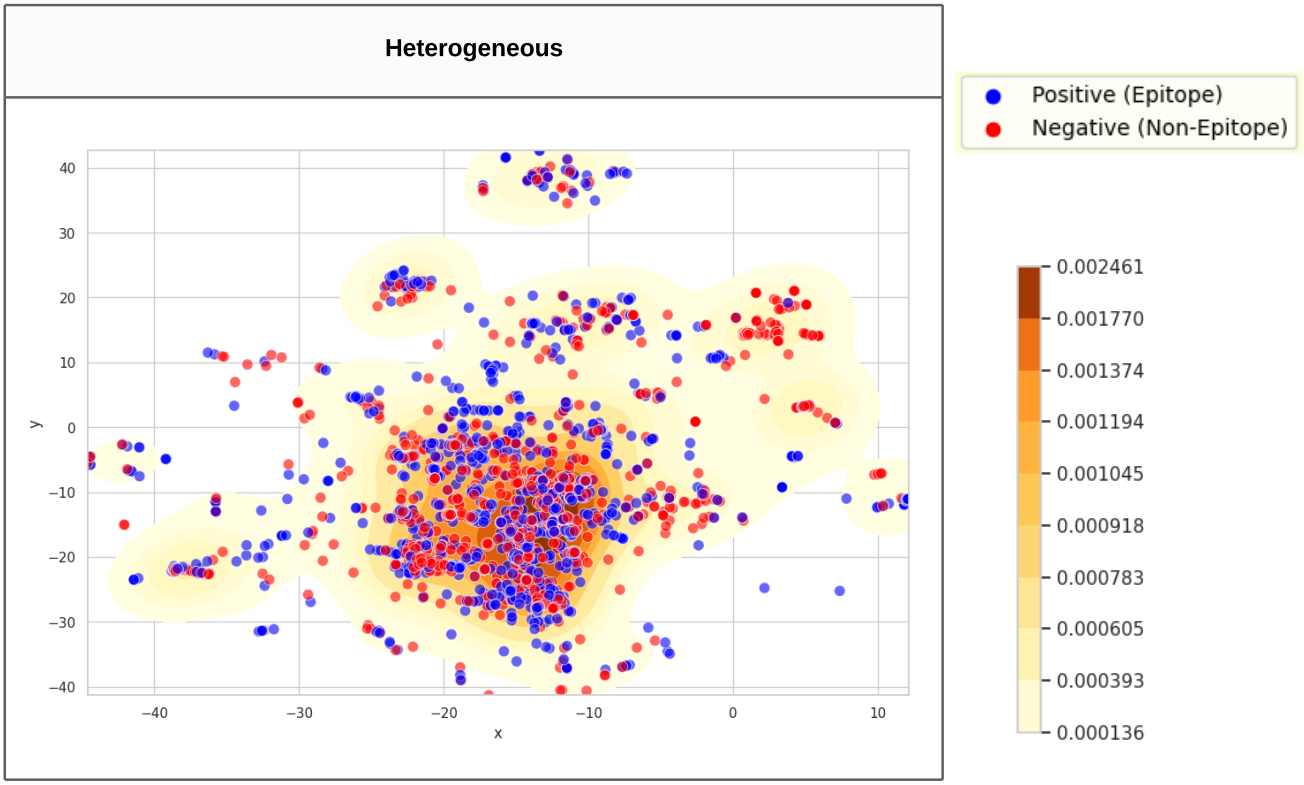

Figure 5.1: The projection was computed using a fraction of the entire dataset. For t-SNE training, 720 samples were employed, and the resulting model was applied to project 2,930 samples (as shown in the figure above) from a larger dataset consisting of 42,990 samples in total. It's worth noting the heterogeneity within the data, consisting of observations from numerous distinct organisms, as positive and negative points appear to be distributed with a certain uniformity across the projected space.

Figure 5.2: The t-SNE projection is presented, stratified by *B. pertussis*, *Corynebacterium and Orthopoxvirus*. Observe the well-defined clusters of high-density positive and negative observations, occupying distinct segments within the feature space. This visual representation illustrates the propensity for epitopes (positive observations) from various pathogens to consistently manifest in separate regions of the feature space. Importantly, regions with a high density of positive examples for one pathogen can also have a high density of negative examples for another pathogen. For instance, the portion around (-20, -5) of the negative *B. pertussis* examples overlaps a high-density region of positive *Corynebacterium* points in the same region. This type of data characteristic may make taxon-specific models better able to learn which regions of the feature space are more strongly associated with positive/negative examples for specific (groups of) pathogens. Generalist models, on the other hand, are trained on data from multiple pathogens, which can make it more difficult for them to learn the specific signatures of each pathogen.

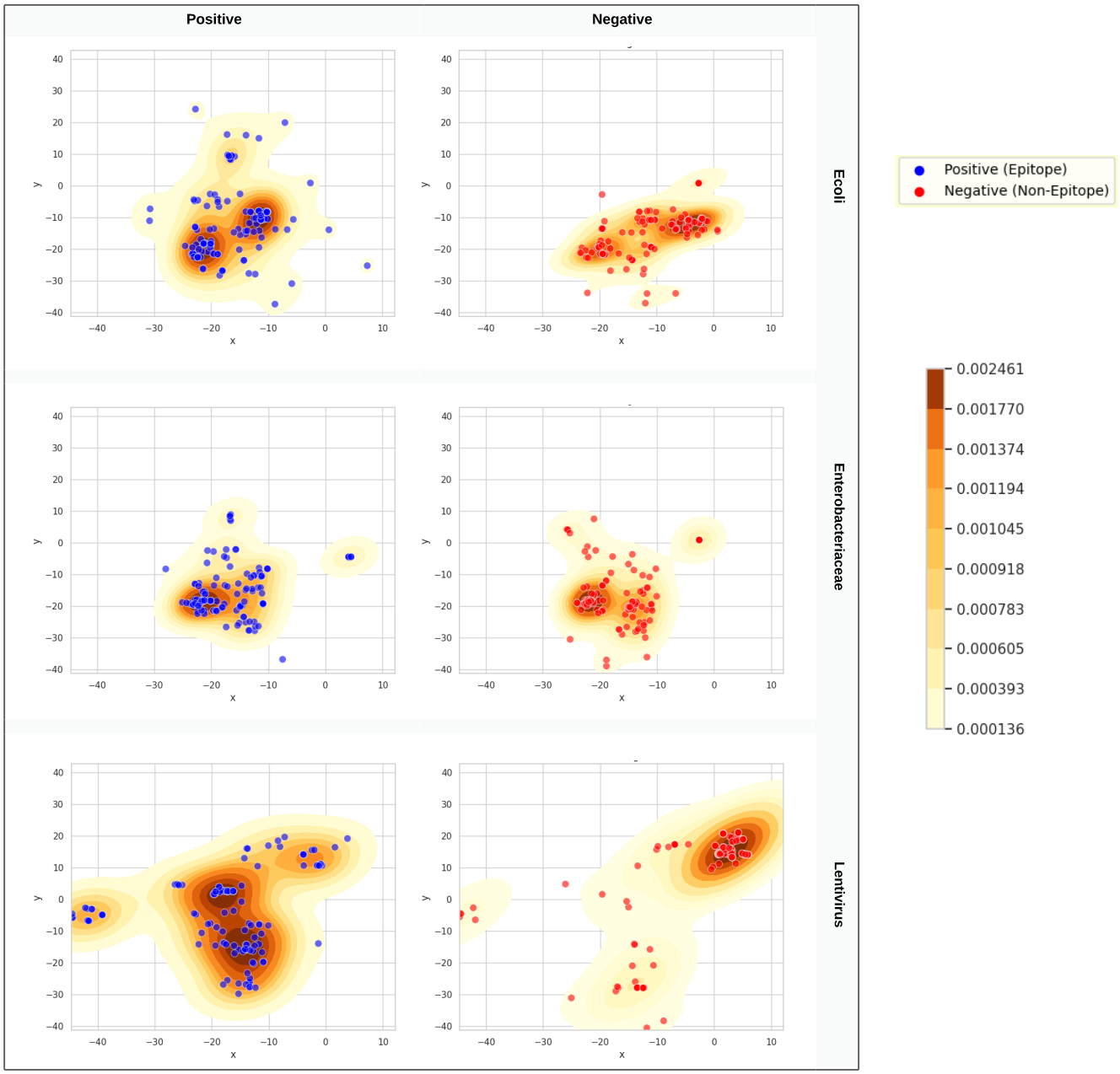

Figure 5.3: t-SNE-projected data from *E. coli, Enterobacteriaceae and Lentivirus*. As another example of why taxon-specific modelling may be preferrable, the negative examples of *Enterobacteriaceae* in the region of (-22, -20) align with a high-density cluster of positive *E. coli* data points in the same region.

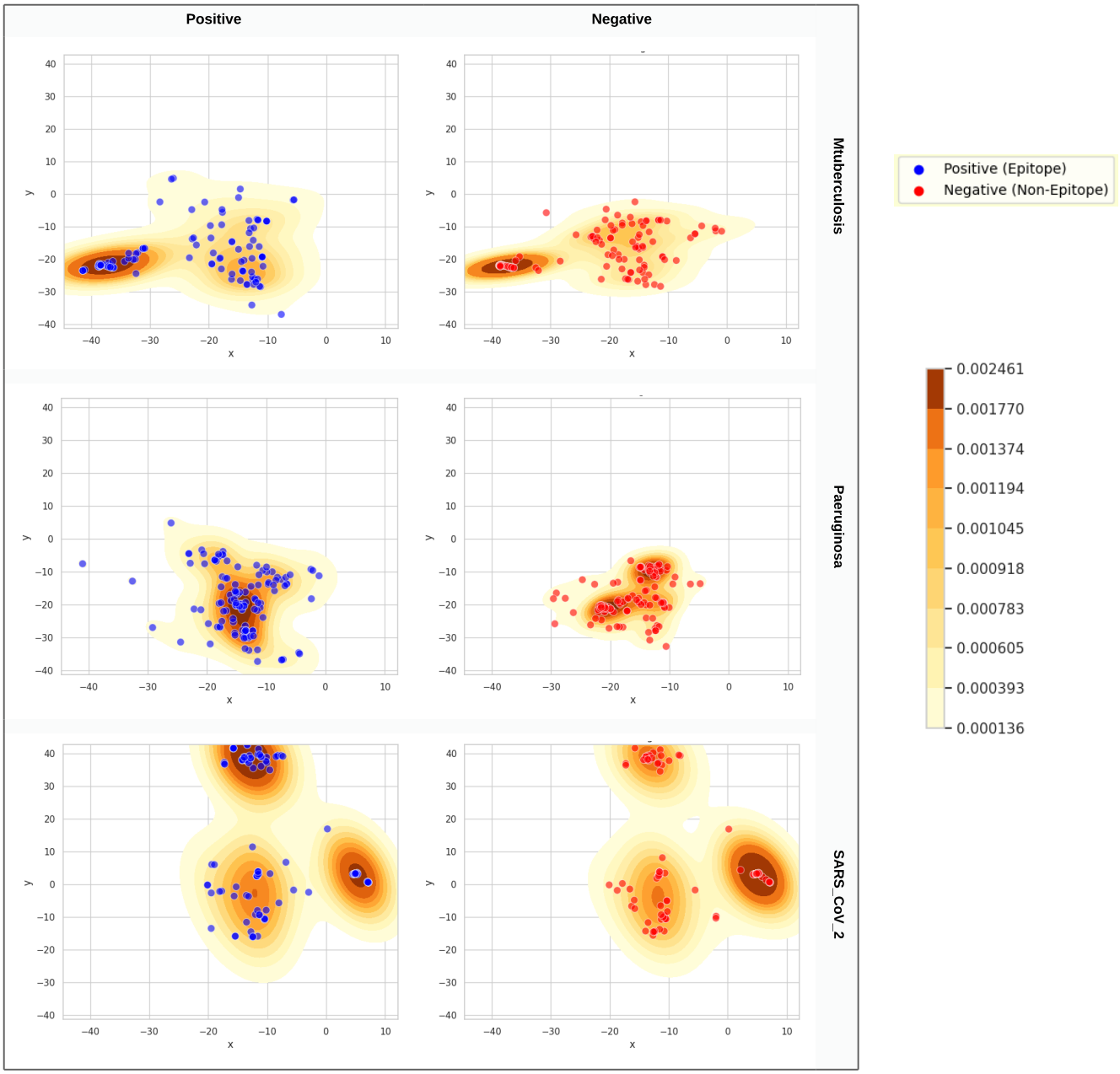

Figure 5.4: t-SNE-projected data from *M. tuberculosis, P. aeruginosa and SARS-Cov-2.* In the projection, the negative examples of *M. tuberculosis*, portion (-15, -20), align with a high-density cluster of positive *P. aeruginosa* data points in the same region. Additionally, note that the negative *P. aeruginosa* samples within the range (-15, -20) roughly coincide with a populated cluster of positive *M. tuberculosis* data points within the same region.

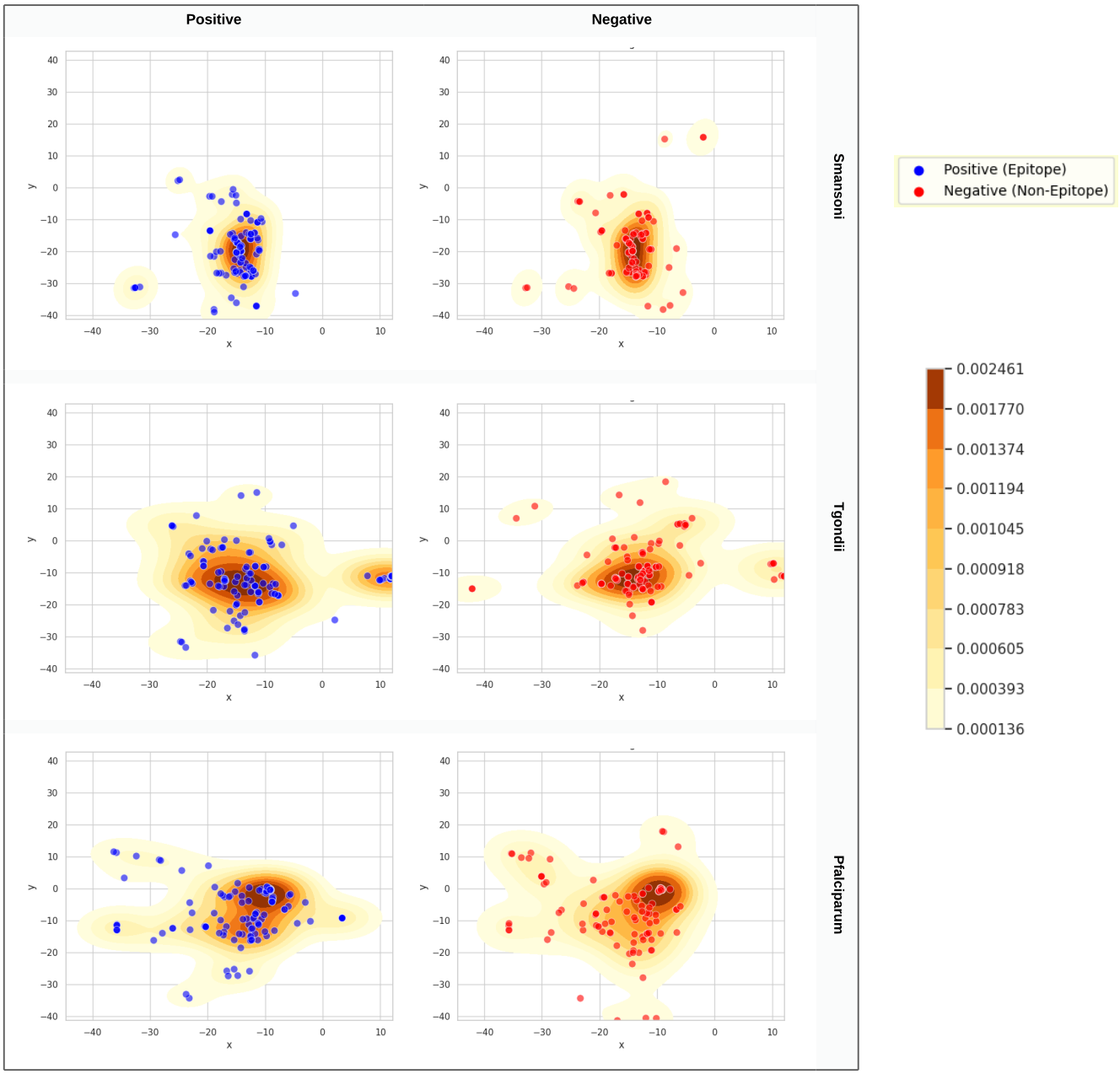

Figure 5.5: t-SNE-projected data from *S. mansoni, T. gondii and P. falciparum.* In this projection, the negative examples of *S. mansoni*, portion (-15, -15), align with a high-density cluster of positive *T. gondii* data points in the same area.

Even though each figure provides only qualitative comparisons for three pathogens, the fact that all figures use consistent coordinates makes it even more significant when creating models to identify overlapping positive and negative regions among different pathogen groups.

5.2 Ablation Study

In the ablation study presented in Table 5.3, the performance of the EpitopeTransfer model is compared to the ESM-1b model, which serves as the base model. ESM-1b is employed in its fundamental configuration, producing embeddings that are fed to a classifier. This architectural approach diverges from that of EpitopeTransfer by excluding the transfer learning step, which involves fine-tuning. The results indicate that EpitopeTransfer consistently outperforms ESM-1b across various organisms and evaluation metrics. Notably, EpitopeTransfer achieves higher AUC, F1, and MCC scores for most of the organisms, presenting its superiority in predicting epitopes. This improvement demonstrates the effectiveness of transfer-learning, where knowledge learned from higher taxonomic levels is transferred to lower taxonomic levels. The results suggest that EpitopeTransfer leverages valuable information from higher-level epitope predictions to enhance its performance in predicting epitopes for specific organisms.

Taxon	Model	AUC	F1	MCC
B. pertussis	esm-1b	0.483 ± 0.039	0.477 ± 0.029	0.227 ± 0.019
	epitope-transfer	0.582 ± 0.030	0.542 ± 0.026	0.263 ± 0.044
Corynebacterium	esm-1b	0.534 ± 0.023	0.540 ± 0.017	0.225 ± 0.036
	epitope-transfer	0.612 ± 0.042	0.600 ± 0.033	0.253 ± 0.063
Orthopox	esm-1b	0.602 ± 0.023	0.583 ± 0.012	0.286 ± 0.035
	epitope-transfer	0.613 ± 0.051	0.591 ± 0.027	0.274 ± 0.025
E. coli	esm-1b	0.849 ± 0.007	0.812 ± 0.013	0.639 ± 0.026
	epitope-transfer	0.924 ± 0.007	0.857 ± 0.013	0.724 ± 0.022
Enterobacteriacea	esm-1b	0.755 ± 0.008	0.723 ± 0.009	0.489 ± 0.011
	epitope-transfer	0.797 ± 0.009	0.767 ± 0.008	0.559 ± 0.014
Lentivirus	esm-1b	0.791 ± 0.005	0.867 ± 0.004	0.764 ± 0.007
	epitope-transfer	0.793 ± 0.006	0.870 ± 0.000	0.770 ± 0.000
M. tuberculosis	esm-1b	0.584 ± 0.004	0.563 ± 0.007	0.176 ± 0.008
	epitope-transfer	0.608 ± 0.004	0.576 ± 0.005	0.196 ± 0.004
P. aeruginosa	esm-1b	0.661 ± 0.012	0.601 ± 0.022	0.382 ± 0.016
	epitope-transfer	0.655 ± 0.048	0.595 ± 0.039	0.390 ± 0.016
SARS-Cov-2	esm-1b	0.557 ± 0.009	0.541 ± 0.010	0.123 ± 0.011
	epitope-transfer	0.576 ± 0.018	0.547 ± 0.018	0.163 ± 0.020
S. mansoni	esm-1b	0.522 ± 0.010	0.532 ± 0.008	0.113 ± 0.011
	epitope-transfer	0.531 ± 0.006	0.540 ± 0.007	0.116 ± 0.014
T. gondii	esm-1b	0.644 ± 0.016	0.616 ± 0.024	0.266 ± 0.030
	epitope-transfer	0.694 ± 0.023	0.636 ± 0.027	0.308 ± 0.045
P. falciparum	esm-1b	0.724 ± 0.005	0.693 ± 0.012	0.472 ± 0.013
	epitope-transfer	0.759 ± 0.006	0.705 ± 0.008	0.465 ± 0.012

Table 5.3: Ablation study comparing the performance of EpitopeTransfer and ESM-1b, which is the base model. The values of AUC, F1, and MCC represent the mean and standard deviation. For each organism, the test dataset was predicted 10 times, and the mean and standard deviation of the predictions were then computed.

EpitopeTransfer outperforms BepiPred 3.0, EpiDope, and EpitopeVec in most cases, except for a few specific metrics on certain pathogens. This is likely due to the fact that some of the test data was used to train the BepiPred 3.0, EpiDope, and EpitopeVec models. Once this data is removed, the performance of these models is expected to decrease.

5.3 Conclusion

The method is evaluated on twelve distinct organism datasets spanning all three domains, showcasing its superiority over baseline techniques. Moreover, t-SNE analysis is employed to offer insights into the improved performance of organism-specific models. The chapter concludes with an ablation analysis, assessing the performance of EpitopeTransfer in contrast to the base model, ESM-1b.

Chapter 6

Conclusion/Work plan

In conclusion, a method is proposed for enhancing linear B-cell epitope prediction through the transfer-learning from higher to lower taxonomic levels using taxon-specific pre-trained models. This technique utilizes the knowledge acquired from higher taxonomic levels to improve prediction performance at lower taxonomic levels. The evaluation of this method is performed on a dataset that includes protein sequences containing experimentally validated linear B-cell epitopes from twelve different organisms or organism groups, across the domains of bacterial, viral, and eukaryotic pathogens. These preliminary results demonstrate that this approach is able to consistently outperform the baseline methods, suggesting it as a valuable and potentially impactful strategy for further investigation.

Based on the demonstrated effectiveness of the proposed method in improving linear B-cell epitope prediction through transfer learning, the final year of the PhD program will focus on the generalization of this approach. This progression is described in Table 6.1, detailing the work plan that includes formalization and generalization of the method, preparation of an academic article, development of additional theoretical chapters, and potential applications beyond epitope prediction.

Research Task	**Due Date**
1 - Formalization and Generalization of the Method	12/2023
2 - Submission of the article to Briefings in Bioinformatics	02/2024
3 - Development of additional chapters on Domain Adaptation, Knowledge Distillation, Feature Selection, and Feature Fusion	06/2024
4 - Application of the method to an additional application	07/2024
5 - Reorganization of the thesis structure	08/2024
6 - Thesis write up finalization and submission	11/2024

Table 6.1: The expanded work plan for the PhD research project.

In Step 1, the focus is on mathematically formalizing the problem in a way that is not restricted to the application domain of epitope prediction, encompassing a broader range of problems where the dataset exhibits a hierarchical structure, in which data elements are interconnected or related. This includes applications such as phylogenetic branching in biological datasets, the evolutionary progression of languages, and the interconnected thematic structures in scientific literature. The completion of this task is aimed for December 2023.

Step 2 involves the final stages of preparing an article for submission to "Briefings in Bioinformatics", a journal with an impact factor of 13.99 and an *Qualis A1* category journal in the CAPES ranking. Currently in the review phase, the article is on track for submission by the scheduled deadline of February 2024.

Step 3 covers the creation of additional chapters focusing on Domain Adaptation, Knowledge Distillation, Feature Selection, and Feature Fusion. These chapters are designed to provide a background, enhancing the understanding of the proposed method and its contributions. The aim is to complete this task by June 2024.

Step 4 focuses on the potential application of the proposed method to a problem beyond epitope prediction, an effort aimed at increasing validation of its effectiveness and adaptability in diverse contexts. It is not a mandatory element of the research, but rather

an extension to illustrate the wider applicability of the method. The expected date for completion of this task is scheduled for July 2024.

Step 5 involves a reorganization of the thesis structure to integrate the proposed changes and ensure that the content is clearly connected. To achieve a more coherent presentation, the thesis will be divided into three parts. Part 1, the Introduction, will include chapters that provide context and background for the thesis. Part 2, Methodology, will contain chapters that describe each research contribution. Part 3, Case Studies, will consist of one chapter for each problem to which the method has been applied, illustrating the practical applications of the research. This restructured format is scheduled for completion by August 2024.

The final stage, Stage 6, involves completing the writing of the thesis followed by a review and critique process by the supervisors, with a target date of November 2024.

References

Altschuh, D., Lesk, A., Bloomer, A., and Klug, A. (1987). Correlation of co-ordinated amino acid substitutions with function in viruses related to tobacco mosaic virus. *Journal of Molecular Biology*, 193(4):693–707. 18

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410. 26

Ashford, J., Cunha, J., Lobo, I., Lobo, F., and Campelo, F. (2021). Organism-specific training improves performance of linear b-cell epitope prediction. *Bioinformatics (Oxford, England)*, 37. 1, 7, 10, 11, 24, 25

Ba, J. L., Kiros, J. R., and Hinton, G. E. (2016). Layer normalization. 13

Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., van Dijk, A. A., Ebrecht, A. C., Opperman, D. J., Sagmeister, T., Buhlheller, C., Pavkov-Keller, T., Rathinaswamy, M. K., Dalwadi, U., Yip, C. K., Burke, J. E., Garcia, K. C., Grishin, N. V., Adams, P. D., Read, R. J., and Baker, D. (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876. 20

Bahai, A., Asgari, E., Mofrad, M. R. K., Kloetgen, A., and McHardy, A. C. (2021). EpitopeVec: linear epitope prediction using deep protein sequence embeddings. *Bioinformatics*, 37(23):4517–4525. 11, 22

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C. (2003). A neural probabilistic language model. *J. Mach. Learn. Res.*, 3(null):1137–1155. 18

Bertoline, L. M. F., Lima, A. N., Krieger, J. E., and Teixeira, S. K. (2023). Before and after alphafold2: An overview of protein structure prediction. *Frontiers in Bioinformatics*, 3. 20

Blythe, M. and Flower, D. (2005). Benchmarking b cell epitope prediction: underperformance of existing methods. *Protein Science*, 14(1):246–248. 10

Brandes, N., Ofer, D., and Linial, M. (2021). Transmembrane protein structure prediction using hierarchical deep learning networks. *Bioinformatics*. 13

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32. 30

Britz, D., Goldie, A., Luong, M.-T., and Le, Q. (2017). Massive exploration of neural machine translation architectures. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1442–1451, Copenhagen, Denmark. Association for Computational Linguistics. 12

Bugnon, L., Fenoy, E., Edera, A., Raad, J., Stegmayer, G., and Milone, D. (2023). Transfer learning: The key to functionally annotate the protein universe. *Patterns*, 4:100691. 23

Campbell, N. A., Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V., and Jackson, R. B. (2021). *Biology*. Pearson, 12th edition. 9

Campelo, F. and Ashford, J. (2022). epitopes: Processing, feature extraction and modelling of epitope data from the immune epitope database (iedb). ″. 26

Campelo, F., Reis-Cunha, J., Ashford, J., Ekárt, A., and Lobo, F. P. (2023). Phylogeny-aware linear b-cell epitope predictor detects candidate targets for specific immune responses to monkeypox virus. *bioRxiv*. 2, 25

Centers for Disease Control and Prevention (2023). Smallpox/monkeypox vaccine information statement. 10

Chandra, A., Tünnermann, L., Löfstedt, T., and Gratz, R. (2023). Transformer-based deep learning for predicting protein properties in the life sciences. *eLife*, 12:e82819. 16

Cheng, J., Bendjama, K., Rittner, K., and Malone, B. (2021). BERTMHC: improved MHC–peptide class II interaction prediction with transformer and multiple instance learning. *Bioinformatics*, 37(22):4172–4179. 12

Chowdhury, R., Bouatta, N., Biswas, S., Rochereau, C., Church, G. M., Sorger, P. K., and AlQuraishi, M. (2021). Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*. 11

Cleveland P. Hickman, J., Roberts, L. S., Larson, A., I'Anson, H., and Eisenhour, D. (2017). *Animal Diversity*. McGraw-Hill Education, 8th edition. 9

Clifford, J., Høie, M. H., Nielsen, M., Deleuran, S., Peters, B., and Marcatili, P. (2022). Bepipred-3.0: Improved b-cell epitope prediction using protein language models. *bioRxiv*. 11, 21, 24

Collatz, M., Mock, F., Barth, E., Hölzer, M., Sachse, K., and Marz, M. (2020). EpiDope: a deep neural network for linear B-cell epitope prediction. *Bioinformatics*, 37(4):448–455. 11, 21, 22

Consortium, T. U. (2022). UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531. 25

da Silva, B. M., Ascher, D. B., and Pires, D. E. V. (2023). epitope1D: accurate taxonomy-aware B-cell linear epitope prediction. *Briefings in Bioinformatics*. bbad114. 9

Dai, Z., Yang*, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-XL: Language modeling with longer-term dependency. 13

Dehghani, M., Gouws, S., Vinyals, O., Uszkoreit, J., and Łukasz Kaiser (2019). Universal transformers. 13

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805. 12, 18, 19

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Yu, W., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., and Rost, B. (2021). Prottrans: Towards cracking the language of lifes code through self-supervised deep learning and high performance computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PP:1–1. 11, 13

Emini, E. A., Hughes, J. V., Perlow, D. S., and Boger, J. (1985). Induction of hepatitis a virus-neutralizing antibody by a virus-specific synthetic peptide. *Journal of Virology*, 55(3):836–839. 10

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 7(2):179–188. 4

Fleri, W., Paul, S., Dhanda, S. K., Mahajan, S., Xu, X., Peters, B., and Sette, A. (2017). The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design. *Frontiers in Immunology*, 8. 6

Forsström, B., Bislawska Axnäs, B., Rockberg, J., Danielsson, H., Bohlin, A., and Uhlen, M. (2015). Dissecting antibodies with regards to linear and conformational epitopes. *PloS one*, 10:e0121673. 8

Fu, L., Zhu, Z., Wu, S., and Li, W. (2012). Cd-hit: Accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)*, 28. 21

Gorbalenya, A. E., Krupovic, M., Mushegian, A. R., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., Simmonds, P., Walker, P. J., Zerbini, F. M., and Kuhn, J. H. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, 5(5):668–674. 9

Gupta, R. S., Lo, B., and Son, J. (2018). Phylogenomics and comparative genomic studies robustly support division of the genus mycobacterium into an emended genus mycobacterium and four novel genera. *Frontiers in Microbiology*, 9. 9

Göbel, U., Sander, C., Schneider, R., and Valencia, A. (1994). Correlated mutations and residue contacts in proteins. *Proteins*, 18(4):309—317. 18

Harris, Z. S. (1954). Distributional structure. 18

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778. 13

Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., and Rost, B. (2019a). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformatics*, 20. 23

Heinzinger, M., Sellner, B., Sturm, M., Unterthiner, T., Garcia, J., and Hochreiter, S. (2019b). Designing neural networks for continuous protein representation learning. *BMC bioinformatics*, 20(1):1–15. 16

Hsu, C., Verkuil, R., Liu, J., Lin, Z., Hie, B., Sercu, T., Lerer, A., and Rives, A. (2022). Learning inverse folding from millions of predicted structures. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S., editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 8946–8970. PMLR. 23

Høie, M. H., Kiehl, E. N., Petersen, B., Nielsen, M., Winther, O., Nielsen, H., Hallgren, J., and Marcatili, P. (2022). NetSurfP-3.0: accurate and fast prediction of protein structural features by protein language models and deep learning. *Nucleic Acids Research*, 50(W1):W510–W515. 21

Jespersen, M., Peters, B., Nielsen, M., and Marcatili, P. (2017). Bepipred-2.0: Improving sequence-based b-cell epitope prediction using conformational epitopes. *Nucleic acids research*, 45. 10, 22

Jiang, H.-W., Li, Y., and Tao, S.-C. (2023). Sars-cov-2 peptides/epitopes for specific and sensitive diagnosis. *Cellular & molecular immunology*, 20(5):540–542. 2

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021a). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589. 16

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S., Ballard, A., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., and Hassabis, D. (2021b). Highly accurate protein structure prediction with alphafold. *Nature*, 596:1–11. 17, 20

Karplus, P. A. and Schulz, G. E. (1985). Prediction of chain flexibility in proteins. *Naturwissenschaften*, 72:212–213. 10

Kimber, A. (1994). An Introduction to the Bootstrap. *Journal of the Royal Statistical Society Series D: The Statistician*, 43(4):600–600. 4

Kitchenham, B. A. and Charters, S. (2007). Guidelines for performing systematic literature reviews in software engineering. Technical Report EBSE 2007-001, Keele University and Durham University Joint Report. 19

Kolaskar, A. and Tongaonkar, P. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. *FEBS Letters*, 276. 10

Kryshtafovych, A., Schwede, T., Topf, M., Fidelis, K., and Moult, J. (2019). Critical assessment of methods of protein structure prediction (casp)-round xiii. *Proteins: Structure, Function, and Bioinformatics*, 87. 17

Larsen, J., Lund, O., and Nielsen, M. (2006a). Improved method for predicting linear b-cell epitopes. immunome res 2:2. *Immunome research*, 2:2. 10

Larsen, J., Lund, O., and Nielsen, M. (2006b). Improved method for predicting linear b-cell epitopes. immunome res 2:2. *Immunome research*, 2:2. 22

Levitt, M. (1976). A simplified representation of protein conformations for rapid simulation of protein folding. *Journal of Molecular Biology*, 104(1):59–107. 10

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2022). Language models of protein sequences at the scale of evolution enable accurate structure prediction. 20

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130. 20

Lin, Z., Feng, M., Santos, C. N. d., Yu, M., Xiang, B., Zhou, B., and Bengio, Y. (2017). A structured self-attentive sentence embedding. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1657–1668. 14

Liu, T., Shi, K., and Li, W. (2020). Deep learning methods improve linear b-cell epitope prediction. *BioData Mining*, 13. 10

Manavalan, B., Govindaraj, R. G., Shin, T. H., Kim, M., and Lee, G. (2018). ibce-el: A new ensemble learning framework for improved linear b-cell epitope prediction. *Frontiers in Immunology*, 9. 10

Margulis, L. and Chapman, M. (2009). *Kingdoms and Domains: An Illustrated Guide to the Phyla of Life on Earth*. Elsevier Science. 8

Matthews, B. (1975). Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA) - Protein Structure*, 405(2):442–451. 4

Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., and Rives, A. (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 29287–29303. Curran Associates, Inc. 23

Merriam-Webster (Accessed: 2023a). Amino acid. 3

Merriam-Webster (Accessed: 2023b). Antigen. 1

Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., Ogunjimi, B., Laukens, K., and Meysman, P. (2020). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, 22(4):bbaa318. 1

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. (2017). Critical assessment of methods of protein structure prediction (casp) - round xii. *Proteins: Structure, Function, and Bioinformatics*, 86 Suppl 1. 17

NCBI (2015). Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 44(D1):D7–D19. 26, 27

Ofer, D. and Linial, M. (2021). Natural language processing of protein sequences: A survey. *Journal of Computational Biology*, 28(1):15–29. 16

Okasha, S. (2019). 63C5Species and classification. In *Philosophy of Biology: A Very Short Introduction*. Oxford University Press. 9

Pandurangan, A. P. and Blundell, T. (2019). Prediction of impacts of mutations on protein structure and interactions: Sdm, a statistical approach, and mcsm, using machine learning. *Protein Science*, 29. 8

Parvizpour, S., Pourseif, M. M., Razmara, J., Rafi, M. A., and Omidi, Y. (2020). Epitope-based vaccine design: a comprehensive overview of bioinformatics approaches. *Drug Discovery Today*, 25(6):1034–1042. 2

Piccaluga, P. P., Di Guardo, A., Lagni, A., Lotti, V., Diani, E., Navari, M., and Gibellini, D. (2022). Covid-19 vaccine: Between myth and truth. *Vaccines*, 10(3). 7

Ponomarenko, J. and Regenmortel, M. V. (2009). B cell epitope prediction. *Structural Bioinformatics*. 1, 6, 7

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. (2019). Language models are unsupervised multitask learners. In ″. 19

Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., et al. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *PNAS*, 118(15):e2016239118. 4, 11, 13, 16, 19, 21

Rodrigues, C. and Plotkin, S. (2020). Impact of vaccines; health, economic and social perspectives. *Frontiers in Microbiology*, 11. 2

Saha, S., Bhasin, M., and Raghava, G. (2005). Bcipep: A database of b-cell epitopes. *BMC genomics*, 6:79. 22

Saha, S. and Raghava, G. (2006). Prediction of continuous b-cell epitopes in an antigen using recurrent neural network. *Proteins*, 65:40–8. 10

Sayers, E. W., Beck, J., Bolton, E. E., Bourexis, D., Brister, J. R., Canese, K., Comeau, D. C., Funk, K., Kim, S., Klimke, W., Marchler-Bauer, A., Landrum, M., Lathrop, S., Lu, Z., Madden, T. L., O'Leary, N., Phan, L., Rangwala, S. H., Schneider, V. A., Skripchenko, Y., Wang, J., Ye, J., Trawick, B. W., Pruitt, K. D., and Sherry, S. T. (2020). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, 49(D1):D10–D17. 25

Shashkova, T. I., Umerenkov, D., Salnikov, M., Strashnov, P. V., Konstantinova, A. V., Lebed, I., Shcherbinin, D. N., Asatryan, M. N., Kardymon, O. L., and Ivanisenko, N. V. (2022). Sema: Antigen b-cell conformational epitope prediction using deep transfer learning. *Frontiers in Immunology*, 13. 7, 23

Shen, W., Cao, Y., Cha, L., Zhang, X., Zhang, W., Ge, K., Li, W., and Zhong, L. (2015). Predicting linear b-cell epitopes using amino acid anchoring pair composition. *BioData Mining*, 8. 10

Singh, H., Ansari, H., and Raghava, G. (2013). Improved method for linear b-cell epitope prediction using antigen's primary sequence. *PloS one*, 8:e62216. 10

Smith, T. F. and Waterman, M. S. (1981). Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197. 26

StudySmarter (2023). Evolutionary fitness: Definition, role and example. 18

UniProt (2020). UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Research*, 49(D1):D480–D489. 26, 27

Van der Maaten, L. and Hinton, G. (2008). Viualizing data using t-sne. *Journal of Machine Learning Research*, 9:2579–2605. 32, 35

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017a). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 12, 15

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017b). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc. 14, 15

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., and et al. (2019). The immune epitope database (iedb): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343. 26, 27

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., and Peters, B. (2018). The Immune Epitope Database (IEDB): 2018 update. *Nucleic Acids Research*, 47(D1):D339–D343. 25

Vita, R., Overton, J. A., Greenbaum, J. A., Ponomarenko, J., Clark, J. D., Cantrell, J. R., Wheeler, D. K., Gabbard, J. L., Hix, D., Sette, A., and Peters, B. (2014). The immune epitope database (IEDB) 3.0. *Nucleic Acids Research*, 43(D1):D405–D412. 25

Vita, R., Peters, B., and Sette, A. (2008). The curation guidelines of the immune epitope database and analysis resource. *Cytometry. Part A : the journal of the International Society for Analytical Cytology*, 73:1066–70. 25

Vita, R., Zarebski, L., Greenbaum, J. A., Emami, H., Hoof, I., Salimi, N., Damle, R., Sette, A., and Peters, B. (2009). The immune epitope database 2.0. *Nucleic Acids Research*, 38:D854 – D862. 22

Väth, P., Münch, M., Raab, C., and Schleif, F.-M. (2022). Proval: A framework for comparison of protein sequence embeddings. *Journal of Computational Mathematics and Data Science*, 3:100044. 13

Wiley (2007). Taxonomic levels. *Encyclopedia of Life Sciences*. 3, 8

Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579. 8

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., Klingner, J., Shah, A., Johnson, M., Liu, X., Łukasz Kaiser, Gouws, S., Kato, Y., Kudo, T., Kazawa, H., Stevens, K., Kurian, G., Patil, N., Wang, W., Young, C., Smith, J., Riesa, J., Rudnick, A., Vinyals, O., Corrado, G., Hughes, M., and Dean, J. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. 12

Yang, X. and Yu, X.-L. (2009). An introduction to epitope prediction methods and software. *Reviews in medical virology*, 19:77–96. 10

Yanofsky, C., Horn, V., and Thorpe, D. (1964). Protein structure relationships revealed by mutational analysis. *Science*, 146(3651):1593–1594. 18

Zhang, R. and Ulery, B. (2018). Synthetic vaccine characterization and design. *Journal of Bionanoscience*, 12:1–11. 7