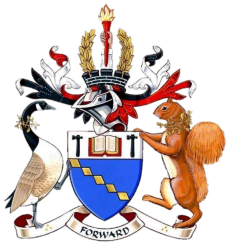
Aston University
College of Engineering and Physical Sciences
Department of Computer Science



PhD Thesis
Doctor of Philosophy in Computer Science

PhD Thesis

Submitted to

Aston University
Department of Computer Science

In partial fulfilment of the requirements for the degree of

Doctor of Philosophy in Computer Science

By

Jodie Sylvia May Ashford

Enhancing Linear B-Cell Epitope Prediction Through Organism-Specific Training

Dr. Felipe Campelo	Supervisor
Prof. Anikó Ekárt	Co-supervisor

September 2023

Author's Declaration of Originality

I hereby certify that I am the sole author of this thesis. All the used materials, references to the literature and the work of others have been referred to. This thesis has not been presented for examination anywhere else.

Author: Jodie Sylvia May Ashford

08-02-2024

Abstract

Aston University
Enhancing Linear B-Cell Epitope Prediction Through Organism-Specific Training
Jodie Sylvia May Ashford
Doctor of Philosophy in Computer Science, 2023

B-cell epitopes play a crucial role in immune responses, with their identification being a vital activity for numerous medical endeavours, including developing diagnostic tests, therapeutic antibodies, and vaccines. Linear B-cell epitopes (LBCE) are often prioritised as targets for epitope predictors over conformational epitopes due to the availability of data, lower experimental complexity for determination and their stability in various conditions, facilitating easier storage and transport. Despite advancements in computational techniques, existing LBCE prediction methods still exhibit suboptimal performance. This thesis explores the efficacy of organism-specific training in improving the accuracy and efficiency of linear B-cell epitope prediction models.

Most LBCE prediction tools adopt a generalist approach, training models on large heterogeneous data sets from numerous organisms to develop predictors that are applicable across a wide variety of pathogens. In contrast, this work investigates the training of bespoke, tailored, organism-specific LBCE prediction models. The main hypothesis posits that using smaller, but potentially more directly relevant, organism-specific data sets for training could yield predictors that demonstrate superior predictive performance for new epitopes of the target organism over a single generalist model.

The main research objectives of this work were: to investigate whether training linear B-cell epitope prediction models using organism-specific data leads to improved prediction performance compared to models trained on heterogeneous or hybrid data, and against well-established epitope predictors from the literature; And to investigate the limits of this organism-specific training approach by systematically quantifying the effect of the amount of training data on the performance of the models developed.

Results indicate that organism-specific training significantly enhances the prediction performance of linear B-cell epitopes, even for organisms with limited training data. Comparative analysis demonstrates the superiority of organism-specific models over heterogenous, hybrid and other conventional predictors, highlighting the potential of tailored modelling approaches in epitope prediction.

Key Words: Epitope Prediction, Machine Learning, Computational Biology

"I may not have gone where I intended to go,
but I think I have ended up where I needed to be."

- Douglas Adams, The Long Dark Tea-Time of the Soul

Acknowledgments

I would like to thank my esteemed PhD supervisor Dr. Felipe Campelo, without whom this endeavor would not have been possible. Felipe's infectious passion for the world of science and research has not only ignited my own enthusiasm but has also been a boundless source of inspiration. I consider myself incredibly fortunate to have had the privilege of learning from him and benefiting from his mentorship throughout this transformative experience.

I would also like to extend my deepest gratitude to Professor Aniko Ekárt, my co-supervisor, whose expertise and guidance have significantly enriched my doctoral journey. Aniko's unwavering commitment to academic excellence and her invaluable insights have been instrumental in elevating the quality of my research.

Thank you Felipe and Aniko, your exceptional academic expertise and commitment to supporting early stage researchers have been priceless assets on my academic journey. I am deeply honored to have had the privilege of learning from two outstanding individuals.

I extend my heartfelt appreciation to Aston University for affording students like myself the invaluable opportunity to pursue a PhD. Thank you for opening the doors to knowledge and enabling us to reach for the highest academic achievements.

Thank you to the UK Research and Innovation Engineering and Physical Sciences Research Council (EPSRC) for funding this work. Their financial support made this research and training possible.

Finally, thank you to my family and friends for their unwavering support and understanding throughout this challenging pursuit.

Table of Contents

List of Figures **8**

List of Tables **9**

General Introduction **10**

0.1 Thesis Structure Overview 11

1 The Epitope Prediction Problem **13**

1.1 Defining Epitopes 13

1.1.1 Linear vs. Conformational Epitopes 14

1.2 Introduction to B-Cell Epitope Identification 15

1.3 Epitope Data Introduction 16

1.4 Defining the Epitope Prediction Problem 17

1.5 Epitope Prediction Tools 17

1.5.1 Linear vs. Conformational Epitope Prediction Methods 19

1.6 Chapter One Conclusion 19

2 Feature Engineering for Epitope Prediction **20**

2.1 Amino Acid Propensity Scales 20

2.2 Machine Learning Feature Sets for Epitope Prediction 21

2.2.1 Machine Learning Feature Sets for Epitope Prediction: Examples 22

2.2.2 Widely Used Features for Computational Epitope Prediction . . . 24

2.3 The Curse of Dimensionality 25

2.4 Optimising Feature Spaces: Dimensionality Reduction and Feature Selection in Epitope Prediction 25

2.4.1 Principal Component Analysis and Kernel PCA 26

2.4.2 Autoencoders 27

2.4.3 Mutual Information 28

2.4.4 Minimum-Redundancy-Maximum-Relevance (MRMR) 28

2.5 The Epitope Dataset 28

2.6 The Epitope Feature Set 30

2.6.1 Representing Protein Sequences - Sliding Window 30

2.6.2 The Feature Set 31

2.7 Chapter Two Conclusion 33

3 Machine Learning Techniques for Epitope Prediction 34
3.1 Trends in Epitope Prediction Machine Learning Techniques 35
3.2 The Epitope Prediction Problem Restated 37
3.2.1 Framing the Machine Learning Problem 37
3.3 Exploring Machine Learning Models for Epitope Prediction 39
3.3.1 Random Forests 39
3.3.2 Support Vector Machines 40
3.3.3 Artificial Neural Networks 40
3.3.4 Gradient Boosting / XGBoost 41
3.4 Evaluating Classifier Performance - Performance Measures 42
3.4.1 Defining Selected Performance Measures 43

4 Organism-Specific Modelling for Linear B-Cell Epitope Prediction 46
4.1 Generalist Modelling Approaches for Epitope Prediction 46
4.2 The Organism Specific Hypothesis 48
4.3 Organism Specific Modelling Pipeline Overview 48
4.4 Organism Specific Dataset Generation 49
4.4.1 Target Pathogens 49
4.4.2 Dataset Generation 50
4.5 Outline of Main Investigations 52
4.6 Modelling 52
4.6.1 Model Selection 52
4.6.2 Hyperparameter Tuning 56
4.6.3 Dimensionality Reduction 57
4.7 Model Testing 59
4.7.1 Performance Assessment and Comparison 60
4.8 Results 61
4.8.1 Organism-Specific Training Improves Performance of Linear B-Cell Epitope Prediction 61
4.8.2 Organism-Specific Models Exhibit Better Performance than Existing Generalist Models 64
4.8.3 Illustrative Example: *Onchocerca volvulus* Results 66
4.9 Insights from Organism-Specific Epitope Prediction Results 71
4.9.1 Exploring Feature Relevance and Implications in the Context of Organism-Specific Modelling 71
4.9.2 Spatial Distribution of Observations 77
4.10 Organism-Specific Modelling for Epitope Prediction Conclusions 78

5 Exploring the Limits of Organism-Specific Training for Linear B-Cell Epitope Prediction 80

5.1 The Scarcity of Epitope Prediction Data 80
5.2 Outlining the Limits of Organism-Specific Training Investigation 81
5.3 The Limits of Organism-Specific Training Investigation Methods 81
5.3.1 Dataset Generation 81
5.3.2 Experimental Protocol Overview 83
5.4 Modelling and Performance Assessment 85
5.5 Results 86
5.6 Discussing the Limits of Organism-Specific Training 89
5.7 Limits of Organism-Specific Training Conclusions 91

6 Discussion 92
6.1 Revisiting Research Questions 92
6.2 Comparison with Existing Literature 93
6.3 Limitations and Future Work 95
6.4 Conclusions 98

References 100

List of Figures

1 Antibody-Antigen Recognition 13
2 Linear vs. Conformational Epitopes 14

3 Protein Sequence Mapping Overview 30
4 Protein Sliding Window 31

5 Classification Task Example 38

6 Epitope Prediction Pipeline Overview 48
7 Model Prediction Summary 59
8 Performance Estimates and Standard Errors of Epitope Predictors 62
9 Aggregated Performance Indices of Organism-Specific Models 63
10 ROC Curves of Epitope Predictors 66
11 Organism-Specific Random Forest Predictions 67
12 Organism-Specific Random Forest Predictions - Part 2 68
13 Feature Importance of Random Forest Models 72
14 Comparison of Feature Importance for Organism-Specific and Heterogeneous Models 72
15 Feature Importance of Organism-Specific Models 73
16 Top Thirty Random Forest Features 75
17 Top Thirty Random Forest Features - Alternate Representation 76
18 Estimated Probability Density of Epitope Observations in the t-SNE Projection 77

19 Limits of Organism-Specific Training Experimental Protocol 84
20 Mean Performance Scores and Standard Errors of Epitope Predictors . . . 86
21 Mean Performance Scores and Standard Errors of Epitope Predictors - Part 2 87

List of Tables

1 Summary of Epitope Prediction Tools 18

2 Overview of Features Extracted from Windowed Sequences 31
3 Amino Acid Types 32

4 Confusion Matrix 42

5 Training Dataset Composition of Epitope Prediction Tools in the Literature 47
6 Number of Examples in each Organism-Specific Dataset 51
7 Default Model Hyperparameters 53
8 Point Estimates of Model Performance 55
9 Random Forest Results After Hyperparameter Tuning 56
10 Default Dimensionality Reduction Parameters 57
11 Random Forest Classifier Performance after Dimensionality Reduction . . 58
12 Performance Estimates and Standard Errors of Epitope Predictors 65
13 Predicted Target Regions on *O. volvulus* Hold-Out Set 69

14 Summary of Organism-Specific Datasets 82

Introduction

An important part of the human immune system's ability to recognise and combat pathogens is dependent on the identification of regions on antigens known as epitopes [1]. Among these, linear B-cell epitopes play a crucial role in triggering the body's adaptive immune response [2]. Accurately predicting the location of linear B-cell epitopes on proteins is a fundamental task in immunoinformatics, with far-reaching implications in vaccine development, disease diagnosis, and antibody therapeutics [3, 4]. Numerous tools exist to predict linear B-cell epitopes from biological data, with a recent emphasis on the development of computational tools and techniques for epitope prediction. These methods leverage various protein features and machine learning techniques to identify potential epitope regions within protein sequences. Although computational techniques have led to significant advancements in epitope prediction, it still remains a complex challenge with considerable scope for improving the accuracy of epitope prediction tools.

In the pursuit of more accurate and reliable linear B-cell epitope prediction, this thesis explores a novel approach: organism-specific training for epitope prediction. Traditionally, many prediction models have been trained on large, diverse datasets that encompass epitopes from a wide-variety of organisms. While these models can provide valuable insights into epitope prediction, they may not fully capture the distinct characteristics of linear B-cell epitopes across different pathogens. The organism-specific training hypothesis is that by tailoring prediction models to the potentially unique epitope patterns of individual organisms, we can significantly enhance prediction accuracy.

This thesis proposes a series of organism-specific training strategies for linear B-cell epitope prediction. Its goal is to investigate the advantages of organism-specific training through a systematic comparison of its performance with conventional heterogeneous prediction models. This work employs an array of machine learning techniques, feature engineering methods, and performance indicators to comprehensively assess this approach. The investigation spans a diverse range of organisms, including viruses, bacteria, and eukaryotes, shedding light on the adaptability and versatility of organism-specific training. The objective is to further the progress of epitope prediction by exploring how organism-specific training can enhance the performance of linear B-cell epitope prediction.

Research Aims and Questions

The main research aims and questions of this thesis are as follows:

Research Aims

- To investigate the potential of organism-specific training in improving the predictive performance of linear B-cell epitope prediction models.
- To compare the effectiveness of organism-specific training with conventional heterogeneous training for epitope prediction.

Research Questions

- Can training linear B-cell epitope prediction models using organism-specific data enhance prediction performance in comparison to models trained on heterogeneous or hybrid data?
- How do organism-specific models compare to well-established epitope predictors from the literature?
- How does the quantity of available organism-specific training data impact prediction performances?
- What is the minimum amount of organism-specific data required for organism-specific models to outperform generalist predictors?

0.1 Thesis Structure Overview

The remainder of this thesis is split into 6 chapters:

Chapter 1 (The Epitope Prediction Problem) introduces the challenge of epitope prediction. It defines what a linear B-cell epitope is; where epitope data may be obtained and presents an overview of epitope prediction techniques.

Chapter 2 (Feature Engineering for Epitope Prediction) will explore the intricacies of feature engineering for epitope data, investigating how features may be extracted from epitope data; highlighting the types of features currently used for epitope prediction; looking at the role of feature selection for epitope feature sets; and finally presenting a carefully crafted feature set designed specifically for the prediction of linear B-cell epitopes.

Chapter 3 (Machine Learning Techniques for Epitope Prediction) provides an exploration of current machine learning (ML) techniques employed in epitope prediction. Additionally, it introduces prevalent performance indicators used to assess epitope predictors and selects the most suitable ones for this study.

Chapter 4 (Organism-Specific Modelling for Linear B-Cell Epitope Prediction) introduces the concept of organism-specific modeling for linear B-cell epitope prediction. It outlines the organism-specific hypothesis; proposes a structured pipeline for organism-specific training; evaluates the models' generalisation performance; conducts comparative analyses with generalist and hybrid models, and with well-established models from the scientific literature.

Chapter 5 (Exploring the Limits of Organism-Specific Training for Linear B-Cell Epitope Prediction) explores the boundaries of organism-specific training in the context of linear B-cell epitope prediction. It assesses the influence of the quantity of organism-specific training data on prediction performance and seeks to identify the minimum amount of organism-specific data required to achieve superior performance compared to models trained on extensive and diverse datasets.

Chapter 6 (Discussion) concludes the thesis, revisiting the research questions posed throughout the study. It also conducts a retrospective comparison between this work and existing approaches from the literature, emphasising the growing importance of more specific training for epitope prediction. Furthermore, this chapter critically discusses the limitations of this work and suggests potential future research avenues that may further advance the field of epitope prediction.

1. The Epitope Prediction Problem

1.1 Defining Epitopes

The immune system exists to protect the body against infection. In many species, including humans, the immune system may be divided into two categories: the innate immune system and the adaptive immune system. The innate immune system, present in the body from birth, gives rise to quick, non-specific immune responses against invading pathogens via proteins, chemicals and cells within the body. The adaptive or acquired immune system, involves pathogen specific responses that take much longer to execute than the fast acting innate immune responses. These specific responses give rise to 'immunological memory'which enables the immune system to respond more rapidly to previously encountered pathogens [1, 2, 5].

Adaptive immunity is vital in the fight against infection with pathogens like bacteria, viruses and parasites. Adaptive immunity can be further categorised into humoral and cell-mediated immunity. Cell-mediated immunity recruits immune cells to destroy infected cells within the body. Humoral immunity involves the production of antigen specific antibodies by B-cells, which are released into the circulatory system to aid in the destruction of extracellular pathogens [2, 6]. This antigen-antibody recognition is a vital process in protecting the body against pathogens and B-cells are key cells in this process.

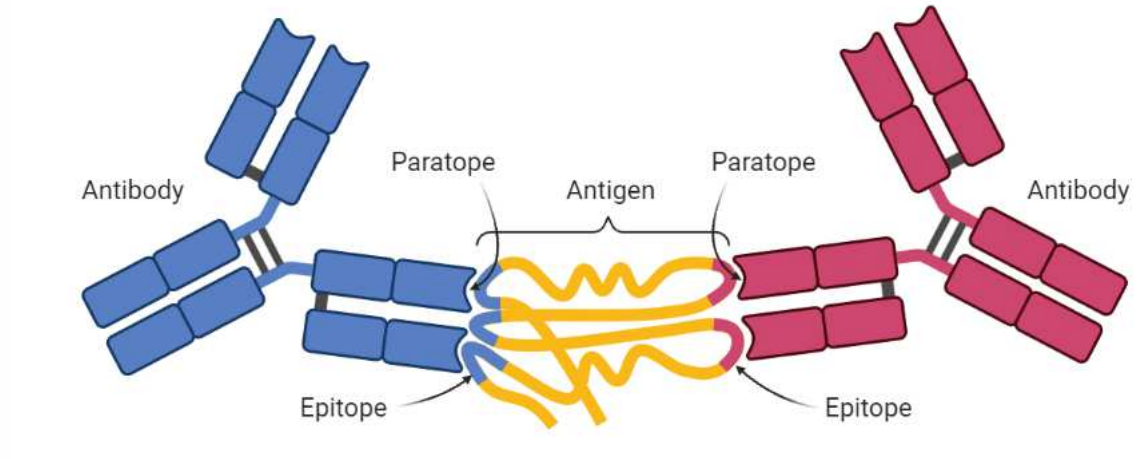


Figure 1. Antibody-Antigen Recognition. Antibodies recognising epitopes on an antigen. Adapted from "Antigen Recognition by Antibodies", by BioRender, August 2020 [7].

An epitope, or antigenic determinant, is the exact portion of an antigen that the antigen-binding site of an antibody recognises and binds to [1, 8]. The portion of the antibody that binds the epitope is known as the antigen-binding site or paratope. Figure 1 shows two different antibodies recognising and binding to two different epitopes on an antigen. B-cell and T-cell receptors (BCR, TCR), are simply membrane-bound antibody molecules present on the outer surface of these cells. A B-cell epitope (BCE) is an epitope that is recognised by a B-cell receptor. Once a BCR recognises its epitope B-cell activation begins. The high specificity of antibody-antigen interactions is key to immunity and can be exploited and used in experimental biology, medical diagnostics, immunotherapies and many other medical and research applications.

1.1.1 Linear vs. Conformational Epitopes

Epitopes can be divided into two categories depending on how their primary amino acid (AA) sequence is recognised by a paratope. The two types of epitopes are linear (or continuous) epitopes and conformational (or discontinuous) epitopes. Figure 2 highlights the difference between linear and conformational epitopes: Linear epitopes are recognised by antibodies by a continuous stretch of amino acids in the proteins primary sequence. Conformational epitopes, on the other hand, are made up of amino acid residues that are separated in the primary sequence but are brought together by protein folding [9][10, Chapter 3], these epitopes are recognised by antibodies by their folded conformation.

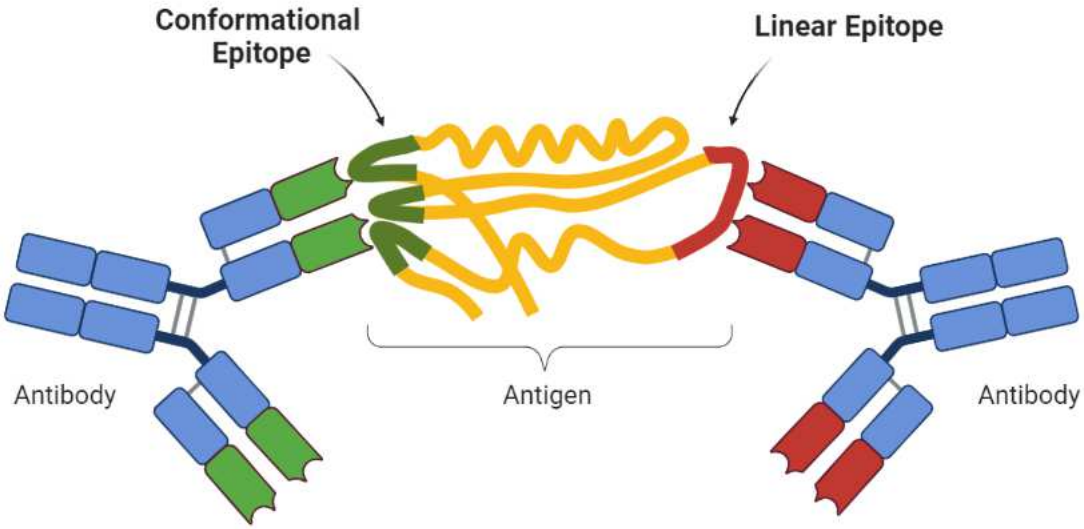


Figure 2. Linear vs. Conformational Epitopes. Conformational epitope shown on the left of the antigen in green, linear epitope on the right in red. Adapted from "Antigen Recognition by Antibodies", by BioRender, August 2020 [7].

Figure 2 shows two different antibodies recognising two different epitopes on the same antigen. The epitope recognised by the antibody on the left-hand side of the diagram is a conformational epitope: here the folded conformation of the protein is important for epitope recognition. The amino acid residues necessary for antigen binding are separated in sequence. The epitope recognised by the antibody on the right-hand side of the diagram is a linear epitope: the AA residues recognised by the antibody are next to each-other in the protein's primary sequence. It has been estimated that the majority of B-cell epitopes are conformational epitopes with approximately $\sim 90\%$ of epitopes on native proteins being conformational BCEs [11–13]. Further details on this are discussed in Section 1.5.1.

1.2 Introduction to B-Cell Epitope Identification

B-cell epitope identification is a crucial process for a number of medical and immunological processes including in vaccine development, therapeutic antibody production and disease prevention and diagnosis [3, 4, 14]. In vaccine development, traditional vaccines, containing live attenuated or inactivated microorganisms, were expensive to manufacture and often carried the risk of creating unwanted immune responses within the patient [4]. Epitope vaccines allow for a more specific and potent immune response [4, 12, 15]. In medical diagnostics epitope arrays have helped to improve the specificity of diagnosis [16], and in therapeutic antibody production epitope prediction has been used to improve antibody quality and utility [17].

Traditionally, experimental methods have been used to identify B-cell epitopes, including: X-ray crystallography, peptide arrays, enzyme-linked immunosorbent assay (ELISA) and phage display [18–20]. These experimental identification methods are often time consuming, resource intensive and technically difficult to execute [14, 18]. In silico epitope prediction methods can significantly reduce the cost of epitope identification over experimental methods. In silico prediction methods use protein sequence or structural data to predict epitope regions on a target. The type of data required for prediction depends on the category of epitope(s) being predicted. Linear epitopes can be predicted from protein sequence data alone whereas conformational epitope prediction usually requires additional 3-Dimensional structural data for prediction. B-cell epitope identification is also highly context-dependent. Whether or not a peptide (short chain of amino acids) is an epitope depends on the host organism (where the immune reaction occurs), the source organism (where the peptide came from) and other specific biological contexts [21]. This context should be taken into account when considering epitope prediction methods.

1.3 Epitope Data Introduction

Since the mid 2000's, access to high throughput DNA sequencing techniques (Next-Generation Sequencing) has greatly accelerated the number and diversity of available protein sequences. Numerous publicly available protein sequence databases currently exist [22–25], many of which are simple protein sequence repositories containing sequences from different sources with little or no storage structure. However, many are expertly curated databases that compile non-redundant sequence data and contain specific metadata for each deposited sequence [25]. In addition to storage of sequence data, many protein structure databases now exist containing numerous experimentally determined or predicted protein structures [26, 27].

Several epitope datasets were curated for this work, the data used for these came from: UniProt [22], the National Center for Biotechnology Information's (NCBI) GenBank and Protein database [23], and the Immune Epitope Database (IEDB) [28]. UniProt (the universal protein knowledgebase) is *"a comprehensive, high quality and freely accessible resource of protein sequence and functional information."* [22], it provides users with high-quality protein sequences annotated with functional information. As of the 12/10/2022 the UniProt Knowledgebase contained 229,928,140 protein sequence entries from 1,291,297 different organisms [29]. The Genetic Sequence Data Bank (GenBank) is the National Institutes of Health's (NIH) genetic sequence database, *"an annotated collection of all publicly available DNA sequences."* [23]. As of 15/10/2022 GenBank contained 240,539,282 nucleotide sequences [30]. The Immune Epitope Database (published by the National Institute of Allergy and Infectious Diseases: NIAID [31]), contains experimental data on both antibody and T-cell epitopes. As of the 04/12/2022 the IEDB contained information relating to 1,551,124 different epitopes [32]. For each of these epitopes key metadata is also provided including, the source organism (following NCBI taxonomy), and the source protein/antigen. Information is also provided regarding the assay used to detect the immune response (e.g. what assay was used and how the response was induced) and on any immune processes associated with the response [21].

The Immune Epitope Database (IEDB) was chosen as the source for epitope data in this thesis as it is one of the most widely-recognised, comprehensive and specialised resources for epitope-related data [33]. It is specifically designed to collect and curate immune epitope data, including labeled linear B-cell epitope data. The IEDB contains high-quality data that has been experimentally verified from a wide variety of pathogens, including viruses, bacteria and eukaryotes (which aligns with the research objective of exploring organism-specific epitope prediction). The data in the IEDB is continually being updated with new literature and data submissions, ensuring that it remains up-to-date with the latest discoveries [34]. The IEDB also provides open access to its data; data is readily available

and easy to obtain, again making it a fitting choice for research on epitope prediction. For this work, epitope data was retrieved from the IEDB and further taxonomic and protein sequence data related to these entries were retrieved from NCBI and UniProt databases. Both NCBI and UniProt are two of the most comprehensive protein databases globally. They provide detailed information, including sequence data, relating to an extensive range of proteins. These resources are known for their high data quality and reliability, curating data from the scientific literature including experimental studies and other trusted sources. Like the IEDB these databases are also continually updated providing access to the most up-to-date protein data. These resources also cross-reference with each-other allowing for IEDB epitope data to be easily linked with its protein data from these sources, making them an ideal choice for accessing protein data for epitope and bioinformatics research.

1.4 Defining the Epitope Prediction Problem

The main focus of this work is to provide a novel organism-specific framework/pipeline for linear B-cell epitope prediction. The epitope prediction problem that this work sets out to solve is: given protein sequence data, predict regions in the data that have a high likely-hood of containing a linear B-cell epitope.

1.5 Epitope Prediction Tools

The first linear epitope prediction method was presented in 1981 by Hopp and Woods [35]. Their method worked by assigning every amino acid residue a hydrophilicity score (numerical value), these values were then averaged along the sequence of the target protein to find the point of greatest local hydrophilicity. An epitope was predicted to be at, or immediately adjacent to, this point [35]. Since 1981, numerous epitope prediction methods have been developed (Table 1). Many of these methods use simple rules like amino acid propensity scales for epitope prediction [9, 36], while others use more complicated techniques like training machine learning algorithms for prediction [37–39].

Generally speaking, epitope prediction methods can be split into two categories: sequence-based prediction methods and structure based prediction methods. Structure-based methods examine the protein's three-dimensional (3D) structure and use this information to make predictions, whereas, sequence-based methods make predictions from sequence data alone. The type of prediction method used is often influenced by the type of epitope being predicted. Sequence-based methods are regularly used to predict linear B-cell epitopes and structure-based methods for conformational epitopes, however, this is not always the case. Though it has been estimated that conformational epitopes make up 90% of all B-cell epitopes on native proteins [13, 40, 41], the majority of current B-cell epitope predictors are designed to predict linear epitopes.

Name / Descriptor	Year	Feature Set	Method / Model
Greatest Hydrophilicity [35]	1981	Hydrophilicity Scores Propensity Scale	Propensity Scale
Hydrophilicity Scale [42]	1986	Hydrophilicity Scale	Propensity scales
Antigenic [36]	1990	Hydrophilicity, accessibility, flexibility, AA counts	Propensity scales
PREDITOP [43]	1993	Hydrophilicity, accessibility, flexibility, secondary structure	Propensity scales
People [44]	1999	Hydrophilicity, accessibility, flexibility, secondary structure	Propensity scales
Bepitope [45]	2003	Accessibility, flexibility, turns, amphiphilicity	Propensity scales
Bcepred [46]	2004	Hydrophilicity, flexibility/mobility, accessibility, polarity, exposed surface, turns	Propensity scales
Söllner/Mayer [37]	2006	AA propensity scales, neighborhood propensities, sequence complexities	kNN and Decision Trees
ABCpred [38]	2006	Amino acid composition	Recurrent Neural Network (RNN)
Chen [47]	2007	AA pair (AAP) antigenicity	SVM
BCPred [48]	2008	AA composition, composition-transition-distribution descriptors, AAP propensity scale	SVM
FBCpred [48]	2008	AA composition, composition-transition-distribution descriptors, AAP propensity scale	SVM
LEPD [49]	2008	Hydrophilicity, flexibility, b-turn, surface accessibility, polarity	Mathematical morphology
Epitopia [50]	2009	Ratio between AA frequencies, polarity, flexibility, antigenicity, hydrophilicity	Naïve Bayes classifier
COBEpro [51]	2009	AA similarities to positive epitope set	SVM
BayesB [52]	2010	Bayes feature extraction-based bi-profile (statistical features)	SVM
LEPS [53]	2011	(2-4) Amino-acid segment features	SVM
BEOracle [54]	2011	Sequence composition, evolutionary information, secondary structure and solvent accessibility	SVM
SVMTriP [55]	2012	Similarity and propensity of tri-peptide subsequences	SVM
BEST [56]	2012	Secondary structure, solvent accessibility, dipeptide-based antigenicity, conservation and similarity scores.	SVM
LBtope [57]	2013	Dipeptide composition	SVM and kNN
BeePro [58]	2013	Physiochemical properties, AA ratio and evolutionary information (PSSM)	SVM
EPMLR [59]	2014	AA composition, hydrophilicity, hydrophobicity, side chain mass	Multiple Linear Regression
DMN-LBE [60]	2015	Dipeptide composition	Deep Maxout Network
LBEEP [61]	2015	Dipeptide deviation from expected mean	AdaBoost-Random Forest
APCpred [62]	2015	AA anchoring pair composition	SVM
Bepipred 2.0 [63]	2017	Computed volume, hydrophobicity, polarity, relative surface accessiblity, secondary structure, overall volume of antigen	Random Forest
DRREP [64]	2017	AA composition (subsequence similarities)	Deep Neural Network
iBCE-EL [39]	2018	AA composition, dipeptide composition, transition-distribution, physiochemical and biochemical AA properties	Ensemble model: extremely randomized tree (ERT) + gradient boosting (GB)
EpiDope [65]	2020	Long Short-Term Memory (LSTM) neural network embeddings	Deep Neural Network
Epitope Vec [66]	2021	AA composition, di-peptide composition, K-mer representation, antigenicity scales & Deep protein sequence embeddings	SVM
BepiPred 3.0 [67]	2022	Protein language model (LM) embeddings	Feed Forward (FFNN), Convolutional (CNN) and Long Short-Term (LSTM) neural networks

Table 1. Summary of Epitope Prediction Tools. Listing the names, publication years, utilised feature sets, and prediction methods of various well-established epitope predictors reported in the literature.

1.5.1 Linear vs. Conformational Epitope Prediction Methods

Linear B-cell epitopes are often predicted from protein sequence data alone whereas conformational epitope prediction usually requires 3-Dimensional structural data. Thanks to recent advances in high throughput DNA sequencing, there is currently a large amount of available protein sequence data. On the other hand, the amount of available real experimentally determined 3D antigen protein structure data is comparably limited (Epitope Data Introduction 1.3) [68]. This is a contributing factor to the popularity of linear B-cell epitope prediction methods over conformational epitope predictors. In silico prediction methods have significantly reduced the cost of epitope identification over experimental methods. However, using 3D structures to predict conformational epitopes is still generally more computationally expensive and time consuming than computational sequence-based linear epitope prediction methods [68]. Additionally, linear epitopes are stable in a wide range of conditions, meaning they are easier to transport and store, making them ideal candidates for peptide vaccines. Conversely, discontinuous epitopes are more easily disrupted by alterations in protein structure caused by factors such as protein-protein interactions and variations in temperature, pH and salinity [69]. Furthermore, it is generally easier to estimate the impact of mutations on linear epitopes, as they are often only affected by amino acid changes in one region, whereas conformational epitopes may be affected by AA changes in many regions of a protein (that result in conformational changes), which are harder to predict and model [70]. For these reasons many epitope prediction studies, including this one, focus on linear B-cell epitope prediction.

1.6 Chapter One Conclusion

Epitope prediction is a crucial step in many medical and immunoinformatic processes. In recent years, the rapid surge in available genomic and proteomic data, coupled with computational advances in machine learning, has revolutionised epitope prediction. This thesis aims to enhance the field of linear B-cell epitope prediction by advocating for a shift towards training epitope predictors on tailored organism-specific training datasets. This work will explore the realm of linear B-cell epitope prediction, navigating through the current state, trends and potential future of the field. The following chapter will explore the intricacies of feature engineering for epitope data and will present a carefully crafted feature set designed specifically for the prediction of linear B-cell epitopes.

2. Feature Engineering for Epitope Prediction

2.1 Amino Acid Propensity Scales

Early epitope prediction methods relied solely on measures of physiochemical features to predict the likelihood of epitopes being present in a given antigen sequence. These features were usually based on properties such as amino acid composition, hydrophobicity, charge, and solvent accessibility, and were used to make simple predictions of linear epitope locations.

The first epitope prediction method by Hopp and Woods (1981) utilised a hydrophobicity score to predict epitope locations in amino acid sequences. In their work, the authors proposed that charged amino acids may be more likely to be part of an epitope as these residues are mainly located on the surfaces of proteins [9, 35]. Other examples of early prediction methods that use physiochemical measures for prediction include works such as: Parker, Guo, and Hodges' hydrophilicity scale (1986) [42]. This method utilised three measures, including surface accessibility, hydrophilicity, and flexibility, to predict B-cell epitopes. A few years later, Kolaskar and Tongaonkar's semi-empirical method (1990) [36] found that specific hydrophobic residues (cysteine, leucine, and valine) were more likely to be present at an antigenic site if located on the protein surface. Using these, along with other physiochemical properties the researchers claimed to predict epitopes with an accuracy of approximately 75 % on a set of 34 proteins. Despite their reported success, these methods are not without their limitations. For instance, Hopp and Woods also noted that not all known epitopes of a protein were located in the most hydrophobic regions [35], highlighting that simple scales may not be able to capture the full complexity of epitope prediction as there are likely many factors that determine epitope location.

Amino acid propensity scales, like the ones mentioned above, are commonly utilised in epitope prediction methods to facilitate accurate prediction. Such scales indicate the likelihood of each amino acid residue to associate with specific properties, such as hydrophilicity and solvent accessibility. These scales work by assigning individual values to each of the 20 amino acid residues based on their relative propensity to possess a specific property. In 1991, Pellequer, Westhof and Regenmortel published a review looking at the effectiveness of different amino acid propensity scales at predicting linear epitopes from primary protein sequences. This study reported prediction accuracies of around 51-57 % for hydrophobicity scales, 46-52 % for accessibility scales and 53-61 % for

β-turn scales [9]. These results demonstrate that none of these scales yield highly accurate epitope predictions. Additional studies have also investigated the performance of these and other simple scale methods on epitope datasets. Saha and Raghava (2004) evaluated the performance of several propensity scales (including [36, 42, 71]) on a dataset of 1029 B-cell epitopes [46]. They reported that the accuracy of the predictions made using these scales varied between 52.92 % and 57.53 % [46]. Blythe and Flower (2005) reported that single amino acid propensity scale methods perform only a marginally better than random at predicting the location of linear epitopes [72], and Kulkarni-Kale, Bhosle and Kolaskar (2005) investigated conformational epitope prediction methods and reported that, for this type of prediction, the accuracy of most propensity scale prediction algorithms lies between 35 % and 75 % [73]. Researchers have also combined multiple propensity scales to create more complex predictors. One such method is Pellequer and Westhof's PREDITOP program, which predicts the location of epitopes using 22 normalised scales corresponding to hydrophobicity, flexibility, accessibility and secondary structures [43]. This study reported an accuracy in correctly predicted epitopes of around 70%. These works indicate that, while amino acid propensity scales are commonly used in epitope prediction methods, relying solely on scale methods based on physiochemical features alone, is not enough to achieve highly accurate predictions.

2.2 Machine Learning Feature Sets for Epitope Prediction

The relatively poor performance of propensity scale methods motivated an increased interest in more sophisticated methods for epitope prediction, such as machine learning (ML) techniques. Machine learning methods are generally considered better than simple propensity scale methods based on physiochemical features alone for epitope prediction [8, 74]. One advantage that these methods have over propensity scale methods is that they can make use of more complex features and patterns that may not be captured by a single or small set of physiochemical properties. Machine learning methods can often handle high-dimensional feature spaces, allowing them to process more data and extract potentially more meaningful patterns than propensity scale methods. They can be trained on large epitope datasets to identify non-linear relationships and dependencies between amino acid residues and their antigenic properties. Machine learning methods are also able to continually learn and adapt to new data, potentially improving their accuracy over time as more epitope data is made available.

Machine learning approaches for epitope prediction typically utilise extensive feature sets, usually extracted from amino acid sequences, to make predictions [37, 39, 54]. The selection of appropriate features for these approaches plays a crucial role in the accuracy of prediction. Features used in machine learning for epitope prediction can be

derived from different sources, the most common ones being statistical or physiochemical properties directly derived from the amino acid sequence of the target protein. Several physiochemical properties can be easily extracted such as the hydrophobicity, size and charge of the individual amino acids that make up the sequence. Numerous other properties can also be calculated using amino acid propensity scales, like solvent accessibility and flexibility, these properties can be used as features for machine learning models. Higher-level sequence patterns can also be extracted from AA sequences, such as motifs (short patterns frequently found in related proteins that often relate to structure or function) and domains (larger structurally and functionally conserved regions of a protein).

In addition to sequence-based features, structural features of proteins are also used for epitope prediction. These attributes may provide information about the conformational and spatial properties of proteins. Examples of structural features used for epitope prediction include: the secondary structure of a protein [75, 76], of which the most common types are the alpha helix (α-helix) and beta sheets (β-sheet). The surface accessibility of residues within a protein can also be calculated from a protein's 3D structure. Surface accessibility scores calculated from sequence data are often predictions based on the physiochemical properties of the amino acid and it's neighbouring residues, whereas those calculated from structure take into account the 3D arrangement of the atoms in the protein. Similarly the solvent accessibility of residues within a protein can also be calculated from the 3D structure.

By using large and diverse feature sets, machine learning models can capture complex patterns in protein data that may not be apparent from a small set of physiochemical properties alone. By leveraging this information, machine learning models are often able to achieve higher accuracy epitope predictions than other prediction methods.

2.2.1 Machine Learning Feature Sets for Epitope Prediction: Examples

Examples of machine learning pipelines for epitope prediction that utilise large feature sets include Söllner and Mayer's *"Machine learning approaches for prediction of linear B-cell epitopes on proteins"* (2006) [37]. This study employed several amino acid propensity scales (including the 55 single amino acid propensity scales from ProtScale [77], and the secondary structure scale proposed by Mayers *et al.,* [78]), and neighborhood descriptors (neighborhood matrices, probabilities & likelihoods and neighborhood complexities) to create a feature set comprising of 18,920 features from each peptide sequence. These features were then utilised along with machine learning algorithms to identify antigenic

determinants (epitopes) on proteins. Feature reduction was performed to generate features sets of between 11 and 164 dimensions. These sets were then passed to 6 different classification models, three k-nearest neighbor (IBk) models and three C4.5 decision trees for classification. These classifiers achieved reported cross-validation accuracies ranging from 62 % to 73 % at classifying peptides as 'epitopes'or 'non-epitopes'.

Another example of a large feature set used for epitope prediction is a study by Wang et al. *"Determinants of antigenicity and specificity in immune response for protein sequences"* / B-Cell Epitope Oracle (BEOracle) (2011) [54]. Here, a feature set consisting of a combination of sequence composition measures, evolutionary information, predicted secondary structure and solvent accessibility measures was used along side a Support Vector Machine (SVM) classifier to predict linear B-cell epitopes. This study utilised a feature set of 53,633 features to train multiple support vector machine models; the highest 5-fold cross-validation prediction accuracy reported was 82.2 % with an F1-measure of 81.37 %. These results were said to outperform classical epitope prediction methods based on propensity [54].

A third machine learning B-cell epitope predictor that utilises a large collection of features extracted from protein sequences is iBCE-EL (2018): *"A new ensemble learning framework for improved linear B-cell epitope prediction"* [39]. This study explored numerous composition measures including: amino acid composition, dipeptide composition, chain-transition-distribution, amino acid index [79] and other physiochemical properties together with six binary profiles as feature sets for ML epitope predictors. The final model (iBCE-EL) is an ensemble predictor that combines two different ML classifiers an Extremely Randomised Tree (ERT) and Gradient Boosting (GB) classifiers to predict linear B-cell epitopes. The reported 5-fold cross-validation accuracies and Matthew's Correlation Coefficients (MCC) of iBCE-EL on a bench-marking and independent dataset are: 73 % accuracy, 0.45 MCC and 73 % accuracy 0.46 MCC respectively.

The examples given above evidence the widespread use of large feature spaces for epitope prediction. As shown, features that are widely extracted from protein sequences include composition measures, physiochemical properties, propensity scales, other statistical properties and occasionally structural features.

2.2.2 Widely Used Features for Computational Epitope Prediction

One set of physiochemical properties that have been used in a number of works are the amino acid index (AAindex) indices [79, 80]. The AAindex is *"a database of numerical indices representing various physiochemical and biochemical properties of amino acids and pairs of amino acids"* [79]. The indices in the AAindex consider both the physiochemical properties of single amino acids and the properties of different amino acid pairs. They have been used for many protein sequence prediction activities [80] including in using SVM's to predict the protein sub-cellular localisation [81], for predicting the immunogenicity of MHC class I binding peptides [82] and in epitope prediction works [39, 72]. Epitope prediction works that have made use of the AAindex include Blythe and Flower's *"Benchmarking B-cell epitope prediction"* study [72] and Manavalan et al.'s ensemble learning framework for B-cell epitope prediction [39].

Another commonly used technique in protein sequence mining is the *n*-gram extraction method. The *n*-gram extraction method was first used by Cherkassky and Vassilas in 1989 for associative database retrieval [83]; it is used to extract contiguous sequences (e.g. multiple consecutive amino acids in sequence) from a given text. The *n* refers to the number of items to be extracted, for example, considering only two consecutive amino acid residues is known as the 2-gram method. Choosing a small value for *n* allows very local patterns and dependencies to be captured from the sequence. This *n*-gram extraction method has been used by several epitope prediction works to extract features from protein sequences [39, 84].

The Conjoint Triad descriptors [85] are another set of feature descriptors that consider consecutive amino acids. These descriptors consider the properties of an amino acid in a sequence together with the properties of the amino acids either side of it; the three continuous amino acids are considered as a unit (triad). Conjoint Triads are based on structural neighbours: amino acids are assigned to one of 7 classes based on their physiochemical properties, resulting in 343 (7^3) possible combinations. The groups are as follows: Group 0 {A, G, V}, Group 1 {C}, Group 2 {F, I, L, P}, Group 3 {M, S, T, Y}, Group 4 {H, N, Q, W}, Group 5 {K, R} and Group 6 {D, E}. These features have been used for many protein sequence mining and prediction activities, including for predicting protein-protein interactions [86], RNA-protein interactions [87], predicting nuclear receptors [88] and in epitope prediction works [20].

2.3 The Curse of Dimensionality

Many epitope prediction pipelines make use of high-dimensional feature sets for epitope prediction. Datasets are often referred to as 'high-dimensional'when the number of features present is significantly higher than the number of entries in the dataset. While utilising large feature sets for epitope prediction may potentially enhance prediction performance, there exist several challenges when employing such high-dimensional feature sets with machine learning algorithms. Large feature sets can pose challenges in terms of storage space and processing time requirements, which can be costly. Storing and managing extensive feature sets may require substantial storage capacity, especially when dealing with high-dimensional data. Additionally, when processing these data, significant computational resources and processing time may be needed. This can result in high infrastructure costs, as hardware or cloud computing resources with greater processing power may be required to handle the data efficiently. Therefore, the expense associated with storing and processing large feature sets should be considered when implementing machine learning pipelines for epitope prediction.

Extensive feature sets may potentially hinder model development in several ways. Large feature sets containing numerous highly correlated features can adversely impact classifier performance [50]. Moreover, feature sets comprising a large number of features and comparatively few data points may lead to overfitting, reducing a model's ability to generalise effectively [89]. The inclusion of too many features can also contribute to overfitting of models due to increasing sparsity of data in high-dimensional feature spaces. This challenge of dealing with an abundance of features in a dataset is often referred to as *'The Curse of Dimensionality'*: a term coined by Richard E. Bellman in 1966 [90]. To mitigate these issues, dimensionality reduction techniques, such as feature extraction and selection, are commonly employed. Many machine learning pipelines for epitope prediction incorporate these techniques to enhance model performance [37, 54, 62].

2.4 Optimising Feature Spaces: Dimensionality Reduction and Feature Selection in Epitope Prediction

Dimensionality reduction is a critical process that involves transforming high-dimensional data into a more manageable and meaningful lower-dimensional representation. This reduction in dimensionality serves several crucial purposes, including conserving storage space, speeding up computational processing, and addressing the challenges posed by the curse of dimensionality, where the performance of machine learning models deteriorates as the number of features increases significantly [91, 92]. The following section delves into various dimensionality reduction techniques, some of which have been used in epitope

prediction pipelines. The section will explore the advantages and disadvantages associated with these methods, shedding light on their role in enhancing the accuracy and efficiency of epitope prediction models.

2.4.1 Principal Component Analysis and Kernel PCA

Principal Component Analysis (PCA) is a linear feature extraction technique first described by Pearson in 1901 [93]. Since then it has been used widely for a variety of data mining applications [94], including computational biology [95]. PCA is a linear transformation that reduces the dimensionality of a dataset while retaining important patterns and trends in the data. This reduction is achieved by projecting the data onto lower dimensions known as principal components (PCs). These principal components represent linear combinations of the original data variables and are arranged in order of decreasing variance [96]. The first PC captures the largest variance, the second PC accounts for as much of the remaining variability and so on [94, Chapter 8.3]. The objective of PCA is to select a limited number of principal components that offer the most informative summary of the data.

PCA, as a feature extraction and dimensionality reduction technique, has found application in various epitope prediction pipelines [20, 37, 97]. For instance: Söllner and Mayer [37] employed PCA alongside other feature reduction methods to reduce a substantial feature set (comprising 18,920 features) into several smaller datasets, each ranging from 11 to 164 dimensions. These reduced datasets were then used as input for k-nearest neighbor and decision tree models for classification. Liu, Yang, and Cheng [97] also demonstrated the effectiveness of PCA as a dimensionality reduction technique in enhancing the prediction quality of B-cell epitopes. While PCA has proven to be a valuable tool for dimensionality reduction in epitope prediction, it's important to note that this method primarily captures linear relationships between features and is sensitive to data scaling.

Traditionally, dimensionality reduction primarily relied on linear techniques, such as Principal Component Analysis (PCA). However, linear methods struggle when dealing with complex, nonlinear data structures. To address this limitation, Kernel Principal Component Analysis (Kernel PCA) emerged as an extension of PCA that harnesses the power of kernel methods to extract principal components. Kernel PCA is a non-linear dimensionality reduction technique that operates in a different space compared to standard PCA and uses kernel methods to find principal components. Unlike standard PCA, kernel PCA is capable of identifying non-linear relationships between features [98]. Though kernel PCA's capability to uncover non-linear data patterns is advantageous, it comes at the cost of increased computational complexity. Nonetheless, this computational investment can be particularly worthwhile when dealing with high-dimensional feature sets, such as those commonly encountered in B-cell epitope prediction tasks.

2.4.2 Autoencoders

Autoencoders [99] are neural networks that can be used for dimensionality reduction of data. Autoencoders compress given input data into a lower dimensional representation and are then able to reconstruct the input from this reduced dimensional representation [94, Chapter 10.4]. Commonly, autoencoders have an encoder-decoder architecture: the encoder maps the input data onto a lower dimensional space and the decoder reconstructs the input from the reduced representation. This hidden reduced 'encoded'representation should represent the most important features from the input data (as these features alone are capable of reconstructing the data). This 'hidden layer'can be extracted from the model and each node can be treated as a new feature similar to the principal components extracted from PCA [94, Chapter 10.4].

Mathematically the encoder stage can be represented as:

$$z = \sigma(Wx + b),$$

where z is the learned reduced data representation, σ is an activation function, W is a weight matrix, x is the input features and b is a bias vector. And the decoder stage can be represented as:

$$\hat{x} = \hat{\sigma}(\hat{W}z + \hat{b}).$$

This maps the encoded representation z to the reconstruction $\hat{x}$. $\hat{x}$ is the reconstruction (the same shape as x) and $\hat{\sigma}$, $\hat{W}$ and $\hat{b}$ are another activation function, weight matrix and bias vector respectively.

Like kernel principal component analysis autoencoders are also capable of modelling complex non-linear functions; they do this by using non-linear activation functions. Again, this process is generally more computationally expensive than standard PCA. Additionally, optimising the weights of non-linear autoencoders is often difficult [100–102] and autoencoders can be prone to overfitting data due to the high number of parameters that they can have. On the other hand, unlike both standard PCA and kernel PCA, the reduced data representation produced by autoencoders retains all of the information from the original dataset. Autoencoders have been used as a form of dimensionality reduction for mass spectrometry imaging, where they are preferred over PCA as they require less human interaction in the analysis [103]. They have also been used for anomaly detection where they have been shown to be able to detect subtle anomalies that linear PCA could not detect [104]. Additionally, autoencoders have been shown to outperform PCA as tasks like image reconstruction [102, 105]. For these reasons, autoencoders could also present a promising avenue for dimensionality reduction for epitope prediction.

2.4.3 Mutual Information

Feature selection is a type of dimensionality reduction that involves selecting a subset of relevant features to use with machine learning systems. Its primary objective is to eliminate 'irrelevant'attributes within datasets, in order to conserve storage space, reduce processing time, and potentially enhance the predictive capabilities of machine learning models [94, Chapter 8.1]. Mutual information is a common measure that is often used for feature selection, it measures the dependency between variables. Two variables are considered independent of each other if they have a score of zero, otherwise, the higher the mutual information score the higher the dependency is between the variables. Similar to dimensionality reduction by kernel PCA or autoencoders, feature selection using mutual information allows non-linear dependencies to be captured from the data [106]. In the context of epitope prediction this feature selection method could prove advantageous for capturing intricate patterns critical for accurate predictions.

2.4.4 Minimum-Redundancy-Maximum-Relevance (MRMR)

Minimum-redundancy-maximum-relevance [107] is a feature selection technique designed to find the smallest relevant subset of features for a machine learning task. It is a minimal-optimal feature selection method. It is often desirable as a feature reduction method as it can help to reduce the memory required and training time to develop ML models and, can help to improve the explainability of the models by only focusing on a small subset of highly relevant features. MRMR has been used as a feature selection method in many bioinformatics contexts, such as in filter-based feature selection for temporal gene expression data [108].

2.5 The Epitope Dataset

Having explored various dimensionality reduction techniques and their potential impact on epitope prediction models, we now turn our attention to the practical aspect of data retrieval. This section outlines the general pipeline for data retrieval used to generate all epitope datasets used in this work. Where methods differ from this routine it will be clearly stated in the text.

All epitope datasets are based on the full Extensible Markup Language (XML) export of the Immune Epitope Database (IEDB) [28]. The database export was filtered according to the following criteria:

- Only peptides marked as linear B-cell epitopes or non-epitopes were selected. The filtering criteria used to isolate peptides marked as linear B-cell entries were:
  - Those with one or more assays containing a *'BCell'* field name in the ***Assay*** fields of the XML export document.
  - Those containing the field *'FragmentOfANaturalSequenceMolecule - LinearSequence'* in the ***EpitopeStructure*** field of the XML export document.

  Peptides marked as either *'Exact Epitope'* or *'Epitope-Containing Region'* in the ***EpitopeStructureDefines*** field were included.
- Only epitopes of lengths between 8 and 25 AA's long were selected. The upper length limit (25) was imposed to prevent overly long sequences labelled as *'Epitope-containing region'* from adding too much noise to the training data. The lower limit (8) was selected to prevent excessive redundancy due to short windows (See below section 2.6.1).
- Record labels *'Positive', 'Positive - High', 'Positive - Intermediate'* and *'Positive - Low'* were grouped under a single label, **'Positive'**.
- Peptide entries with multiple assay results that had conflicting class labels (Positive versus Negative/Non-Epitope) were assigned a class label determined by simple majority. Ties were removed from training datasets.

Each record in the IEDB contains a protein ID, referencing the protein that the peptide entry is a part of/belongs to. For each entry, these IDs were retrieved and used to query NCBI's protein database [109] and UniProt [110]. The entire protein sequence (of which the record peptide belongs to) was then linked to the record. Observations with missing, invalid or inconsistent protein identification information (protein ID or peptide position on the protein) were removed.

Figure 3 shows an overview of the sequence mapping process performed for each filtered IEDB entry. The databases (UniProt & NCBI) are queried using the protein ID retrieved from each IEDB entry. Once the full protein sequence is recovered from a database, the entry peptide (sub-sequence) can then be mapped onto the full protein sequence. Mapping the labelled peptide onto it's full protein sequence is useful as it allows information to be extracted regarding the surrounding local chemical environment of the peptide. Once each entry in a curated dataset has been mapped onto it's full protein, the entries are then windowed and several features are extracted from each peptide window to form the epitope feature set.

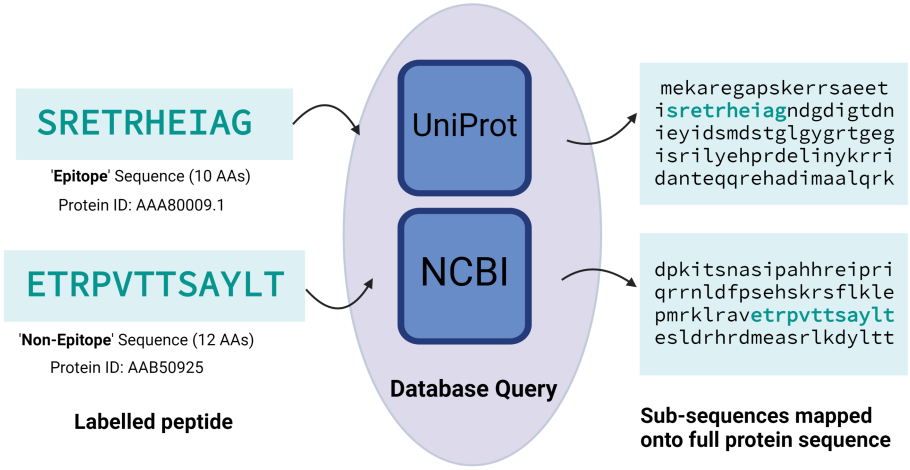


Figure 3. Protein Sequence Mapping Overview. Full protein sequence retrieved from UniProt or NCBI and positions of target peptide on the full sequence are recorded. Created in BioRender, March 2023 [7].

2.6 The Epitope Feature Set

This work makes use of a relatively large feature set comprised of numerous features calculated from the amino acid sequence alone. 845 simple features are calculated for each amino acid residue in a given peptide, based on the local neighbourhood of every position extracted using a 15-AA sliding window representation with a step size of one.

2.6.1 Representing Protein Sequences - Sliding Window

In this work a fixed-width sliding window (15 AAs long) is applied over every peptide entry, with a step size of one. The technique of windowing protein sequences is often used in epitope prediction pipelines [9, 38, 45, 56, 64, 66, 111] as it enables the extraction of relevant local features from micro-environments within the protein sequence, which can be informative for predicting the location of epitopes in the sequence. This is especially important for linear epitopes, which are typically composed of a short stretch of amino acids within a larger protein sequence. By sliding windows of a specific size across the protein sequence, relevant information regarding the local amino acid composition and physicochemical properties of the protein can be extracted.

As mentioned, in this work every record in a collated epitope dataset if first mapped onto the protein sequence that it comes from (Figure 3). This allows for each target residue, in a labelled peptide entry, to be positioned in the centre of the sliding window; enabling the local chemical neighbourhood of that residue to be captured. Figure 4 shows a fixed-width sliding window, size 15 AAs long, being moved across a peptide with a step size of 1.

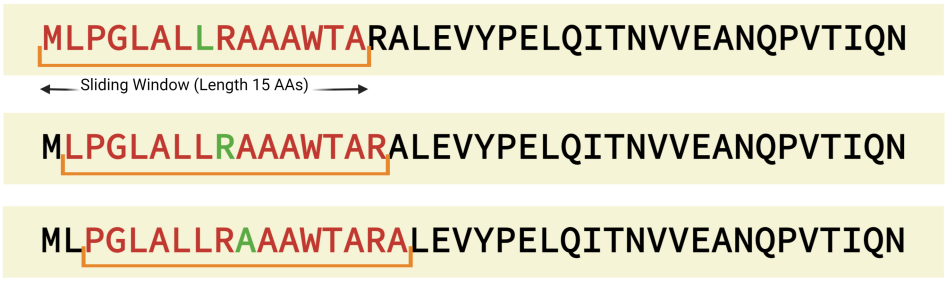


Figure 4. Protein Sliding Window. Fixed-width sliding window, 15-AA long, is applied over the target peptide, with the target residue positioned in the centre of the window. Created in BioRender, February 2023 [7].

For each peptide, length n, the resulting windowed dataset will be comprised of exactly n windows, with each window becoming a single example in the new dataset. The choice of window length 15 was based on the smallest peptide length of interest of the epitope datasets being investigated, namely 8. A window length of 15 AAs therefore provided the longest possible window such that more than half of the residues covered by the window would belong to a labelled (epitope or non-epitope) peptide.

2.6.2 The Feature Set

The following features were then calculated for each window/sample in a given dataset:

Feature(s)	**Number**
Proportion of Amino Acid Residues	20
Conjoint Triad Frequencies	343
Frequency of Amino Acid Types	9
K-mer Composition	400
Number of Atoms	5
Molecular Weight	1
Sequence Entropy	1
Amino Acid Descriptors	66

Table 2. Overview of features extracted from each window sequence. Categories of features extracted from each amino acid window sequence alongside the number of features extracted for each category.

- **Proportion of individual amino acids in the sequence** (20 features): The frequency /percentage composition of each amino acid residue in the sequence.
- **Conjoint Triad frequencies** [85] (343 features): The 20 standard amino acids can be clustered into seven classes according to their dipoles and volumes of their side chains ({A,G,V}, {I,L,F,P}, {Y,M,T,S}, {H,N,Q,W}, {R,K}, {D,E}, {C}). Any three adjacent amino acids in a sequence are regarded as a unit (a triad). Triads are then categorised according to the classes of amino acids, i.e. triads composed of three AAs belonging to the same classes are treated as identical, for example

AIY and GLM. Using the 7 category reduced representation, the frequency of each triad type is counted. Each protein sequence can then be represented by a 7x7x7 dimensional vector (343 features).

- **Frequency of amino acid types** (9 features) [112]: The 20 standard amino acids are grouped based on the properties of their side chains, their size, hydrophobicity, charge and response to pH 7 (Table 3). The frequency of each AA type in the sequence is then calculated.

Class	**Amino Acids**
Tiny	A, C, G, S, T
Small	A, B, C, D, G, N, P, S, T, V
Aliphatic	A, I, L, V
Aromatic	F, H, W, Y
Non-Polar	A, C, F, G, I, L, M, P, V, W, Y
Polar	D, E, H, K, N, Q, R, S, T, Z
Charged	B, D, E, H, K, R, Z
Basic	H, K, R
Acidic	B, D, E, Z

Table 3. Amino Acid Types. 20 standard AA's grouped by side-chain properties.

- ***K*-mer composition** (400 features): A k-mer is a sub-sequence of length k of the amino acid sequence. E.g. the sequence ATGK is made up of four monomers (A, T, G, K) and three 2-mers (AT, TG, GK). For this study, $k = 2$. The frequency of each k-mer within the sequence is calculated. When $k = 2$ for any given sequence there are 400 possible 2-mer combinations (20x20).
- **Number of atoms** (5 features): The total number of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms in the sequence.
- **Molecular weight** (1 feature): The total molecular weight of the sequence.
- **Sequence Entropy** (1 feature) [113]: The information entropy of the distribution of amino acid residues in the sequence.
- **Amino Acid Descriptors** (66 features): From the R *'Peptides'* package, [114] aaDescriptors computes 66 descriptors for each amino acid of a protein. These descriptors are:
  - **Cruciani properties** (3 features) [115]: Three scales that characterise side chains according to their polarity (PP1), hydrophobicity (PP2) and H-bonding capability (PP3).
  - **Kidera factors** (10 features) [116]: Helix/bend preference (KF1), side-chain size (KF2), extended structure preference (KF3), hydrophobicity (KF4), double-bend preference (KF5), partial specific volume (KF6), flat extended preference (KF7), occurrence in alpha region (KF8), pK-C (KF9) and surrounding hydrophobicity (KF10).

- **Z-scales** (5 features) [117]: The computed average of Z-scales of all the amino acids in a given sequence. Each z-scale represents an amino acid property: Lipophilicity (Z1), steric properties (steric bulk/polarisability) (Z2), electronic properties (polarity/charge) (Z3), relating electronegativity, heat of formation, electrophilicity and hardness (Z4 and Z5).
- **Factor Analysis Scales of Generalised Amino Acid Information (FASGAI) vectors** (6 features) [118]: The computed average of FASGAI indices of all the amino acids in a given protein sequence. Each factor represents an amino acid property: Hydrophobicity index (F1), alpha and turn propensities (F2), bulky properties (F3), compositional characteristic index (F4), local flexibility (F5) and electronic properties (F6).
- **T-scales** (5 features) [119]: The T-scales are derived from principal component analysis (PCA) of 67 common structural and topological descriptors of amino acids.
- **VHSE-scales** (8 features) [120]: The VHSE-scales are principal component score Vectors of Hydrophobic, Steric and Electronic properties, derived from PCA on 18 hydrophobic properties, 17 steric properties and 15 electronic properties. Each scale represents an amino-acid property: hydrophobic properties (VHSE1 and VHSE2), steric properties (VHSE3 and VHSE4) and electronic properties (VHSE5, VHSE6, VHSE7 and VHSE8).
- **ProtFP descriptors** (8 features) [121]: These descriptors were constructed from a large selection of indices from the AAindedx database [79].
- **ST-scales** (8 features) [122]: The ST-scales take 827 constitutional, topological, geometrical, hydrophobic, electronic and steric properties into account.
- **BLOSUM indices** (10 features) [123]: The BLOSUM indices were derived from physiochemical properties subjected to VARIMAX analyses and an alignment matrix of the 20 natural AAs using the BLOSUM62 matrix.
- **MS-WHIM scores** (3 features) [124]: These scores were derived from 36 electrostatic properties derived from the three-dimensional structure of the 20 natural amino acids.

2.7 Chapter Two Conclusion

Chapter 2 has outlined the landscape of feature engineering for epitope prediction. It has explored ways in which features can be extracted from epitope data and shed light on the types of features currently prevalent in the field. Feature selection and engineering are pivotal steps in enhancing predictive accuracy for epitope predictors; presented here is a meticulously crafted feature set designed explicitly for predicting linear B-cell epitopes. While this full feature set is likely to contain some redundant features, this issue will be further explored in later chapters that fully describe the full epitope prediction pipeline.

3. Machine Learning Techniques for Epitope Prediction

Machine learning is the foundation of modern predictive modelling, it entails the development of algorithms and models capable of learning from data to make predictions. The process of utilising ML algorithms to build models from data is known as *training*, and the data used to build these models is called the *training data*. Once a model has been trained, the process of making predictions with a learned model is called *testing* and the data to be predicted on is called the *test data* [125]. One of the fundamental goals of machine learning is to develop models with high *generalisation* power, which is the effectiveness of the models on unseen test data. High generalisation performance is pivotal as it ensures that the ML models can accurately predict outcomes on new, unseen test data. When ML models perform well on the training data but have poor generalisation performance, they may have overfit to the training data set. *Overfitting* is where the ML model excessively learns from the training data, resulting in diminished performance on the test data [126]. Models that cannot make accurate predictions on new data are said to have poor generalisation performance. Ensuring that models can effectively predict outcomes on new, unseen test data is vital for computational epitope prediction. The following chapter explores current machine learning techniques utilised in epitope prediction, aiming to contextualise the epitope prediction challenge within the context of machine learning problems.

In recent years, the field of epitope prediction has witnessed a significant shift towards using machine learning techniques for prediction. Machine learning methods, including but not limited to support vector machines [127], random forests [128], and neural networks [129, 130], have become standard tools in epitope prediction. An inherent advantage of employing machine learning techniques like these is their capability to unearth intricate patterns and discern non-linear relationships within complex epitope datasets. Unlike traditional rule-based or statistical approaches, machine learning algorithms excel at capturing dependencies between the large feature spaces associated with epitope data. This allows these models to discern subtle, biologically relevant signals that might otherwise remain hidden. Furthermore, machine learning techniques like these often excel at managing high-dimensional datasets, a frequent characteristic of epitope prediction problems. Some models can efficiently navigate extensive feature spaces, while others can autonomously identify the most informative features, reducing dimensionality without sacrificing predictive power. Additionally, machine learning pipelines are adaptive and able to continuously improve their predictive accuracy as new data becomes available. When new experimental epitope data emerges, machine learning models can be seamlessly retrained on larger and

more diverse datasets, potentially leading to enhanced prediction performance. These models can also undergo optimisation and fine-tuning further improving their performance. These are a few of the reasons why machine learning algorithms have become a popular choice for epitope prediction.

The remainder of this chapter explores the current trends in machine learning techniques for epitope prediction, specifically linear B-cell epitope prediction, examining the advantages and disadvantages of these methods. Furthermore, this chapter aims to conceptualise the epitope prediction challenge within the context of a machine learning problem. Additionally, it outlines a range of machine learning models chosen for investigation in this study and presents the performance indicators selected for evaluating epitope predictors in this work.

3.1 Trends in Epitope Prediction Machine Learning Techniques

Current epitope prediction works employ a range of machine learning (ML) techniques. ML epitope predictors also exhibit a wide spectrum of training approaches; some rely solely on sequence data, others incorporate 3D protein structures, and some combine features from propensity scales and various other sources. Machine learning methods for epitope prediction generally outperform those based solely on simple amino acid propensity scale calculations [8], although there are exceptions to this trend [8, 74]. Examples of machine learning approaches for epitope prediction include: neural network-based methods such as ABCpred [38], Support Vector Machines [127, 131] which have been used in many epitope prediction pipelines [47, 48, 51, 52, 54, 56–58, 62, 132] and Random Forest Classifiers [128, 133] which have also been used in multiple epitope prediction pipelines [61, 63, 134].

As discussed in Section 2.1, early approaches to epitope prediction were characterised by rudimentary rule-based methods and amino acid propensity scales (Section 2.1, Table 1) [35, 36, 42–46]. These approaches rely on simple heuristics and expert knowledge to identify potential epitopes. However, they are limited and unable to capture complex relationships and non-linear patterns within epitope data. The relatively poor performance of these predictors [72], coupled with the sudden availability of extensive sequence data resulting from next-generation sequencing, facilitated the transition toward more sophisticated methods, particularly machine learning methods, for epitope prediction.

Machine learning models such as Support Vector Machines (SVMs), Decision Trees, and Neural Networks, emerged as powerful tools capable of capturing intricate patterns and non-linear relationships within epitope datasets. From 2005 to 2015, Support Vector Machines held a position of prominence as popular models for epitope prediction. During this period, examples of SVM predictors included BCPred, developed by Manzalawy, Dobbs and Honavar, as described in the study *"Predicting linear B-cell epitopes using string kernels"* (2008) [48]. This work evaluated SVM classifiers employing five different kernel methods. The evaluation was conducted using a dataset comprising 701 linear B-cell epitopes (sourced from the Bcipep database [135]) and an equal number of non-epitopes (from SwissProt sequences [22]). The reported predictive performance of BCPred was an AUC score of 0.758, supposedly outperforming eleven other SVM-based classifiers from the literature. Another example of a SVM epitope prediction approach from this period was SVMTriP: *"A method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity"* by Yoa, Zhang, Liang and Zhang (2012) [132]. SVMTriP integrated tri-peptide similarity and propensity scores to train a SVM classifier. This approach demonstrated reported predictive performance scores of 80.1% sensitivity, 55.2% precision, and an AUC score of 0.702 when assessed through five-fold cross-validation. These results showcased the potential of SVM-based methods in epitope prediction during this period.

Though widely-used (Table 1), support vector machines were not the only ML methods emerging for epitope prediction during this time. The first epitope prediction method to use neural networks, ABCpred [38], was published in 2006. In this study, Saha and Raghava trained a recurrent neural network (RNN) to predict linear B-cell epitopes from antigen sequences. They curated a non-redundant dataset of 700 continuous B-cell epitopes (obtained from the Bcipep database [135]), and 700 peptides estimated not to represent epitopes (obtained randomly from UniProtKB [110]). This dataset was then used to train and test several standard feed-forward (FNN) and recurrent neural networks (RNN) at predicting B-cell epitopes on antigens. Their best performing model was a RNN (with a single hidden layer), it yielded a reported 66 % five-fold cross-validation accuracy, 67 % sensitivity, 65 % specificity and 66 % positive predictive value. In addition to SVMs and NNs, other ML models being employed for epitope prediction at this time included Decision Trees [37], Naive Bayes classifiers [50] and K-Nearest Neighbour classifiers [37, 57]. While these models demonstrated the capacity to uncover intricate patterns and non-linear relationships within epitope datasets, surpassing the capabilities of earlier rule-based approaches, the need for even higher prediction accuracies in epitope prediction models persisted.

Post-2015, there was a notable shift in the field of epitope prediction towards the increased use of deep learning models, hybrid models, and ensemble learning methods. These approaches gained popularity due to their ability to capture complex patterns in epitope data and potentially improve predictive accuracy. Since 2015, machine learning techniques including Deep Maxout Networks [60], Random Forests [63], AdaBoost Random Forests [61], ensemble models using Extremely Randomised Trees and Gradient Boosting [39], and deep-learning techniques [64, 65, 67] have been used for epitope prediction. These studies often report significantly improved prediction performances when compared to previous ML methods. In addition to the increased use of deep-learning models for prediction, there's been a recent surge in leveraging large protein language models for feature representations for training classifiers for epitope prediction [65–67]. While deep-learning methods offer significant potential in advancing epitope prediction, they do come with certain drawbacks. These include challenges related to interpretability, susceptibility to overfitting training data, and being resource-intensive and expensive to use. Overall, these advancements reflect the growing recognition of the importance of accurate epitope prediction in vaccine design and immunoinformatics. Researchers are increasingly exploring more complex and sophisticated modeling techniques to enhance the predictive performance of epitope prediction models.

3.2 The Epitope Prediction Problem Restated

Previously, the epitope prediction problem at the core of this research was defined as follows: *Given protein sequence data, predict regions in the data that have a high likelyhood of containing a linear B-cell epitope.* One of the first major challenges was to define and frame this problem as a machine learning problem. This section attempts to establish this epitope prediction challenge as a machine learning problem.

3.2.1 Framing the Machine Learning Problem

The problem at hand is the accurate prediction of where linear B-cell epitopes are within protein sequences. This can be distilled to: for each amino acid residue within a given protein sequence, can it be classified as belonging to an epitope or not? Machine learning models are algorithms that are able to learn patterns and dependencies from provided data. The issue of predicting where an epitope may be on an unlabelled peptide can be presented as a supervised classification task for a ML model. As introduced in section 1.3, there are several curated databases containing labelled epitope and non-epitope data. Labelled data can be used to train ML models to make predictions on unseen/unlabelled data. In machine learning, supervised learning is where a ML algorithm learns from a labelled training dataset, to make predictions on new unseen data [136]. Unsupervised learning, on the other hand, is where models learn from unlabelled data and includes tasks like clustering

and dimensionality reduction [125]. Supervised learning is often further broken down into two categories: *Classification* tasks and *Regression* tasks. In classification problems, the target class label, that the algorithm is trying to predict, is categorical (falls into a distinct category), whereas, in regression tasks the variable that the algorithm is trying to predict is continuous or numeric. The epitope prediction problem outlined here is a classification task. The task is to predict one of two potential class labels: either *'epitope'* if the residue in question is thought to belong to a linear B-cell epitope or *'non-epitope'* if the residue does not belong to an epitope.

Training datasets for this classification task require numerous labelled peptides from each class ('epitope'or 'non-epitope'). Linear B-cell epitope predictors are often sequence-based predictors that make use of labelled sequence data to make predictions. This work makes use of labelled protein sequence data (from the IEDB) and employs a sliding window technique as a pre-processing step, to represent the data in a format that is more easily amenable to the calculation of features and the downstream prediction of arbitrary-length epitopes. As discussed in section 2.6.1, sliding window techniques are frequently employed in epitope prediction; this approach is used as linear B-cell epitopes vary in length, and the exact location of unknown epitopes within a protein is uncertain. Sliding windows also enable the extraction of relevant local features from protein data. Several physiochemical and statistical features are extracted from the windowed dataset entries, this feature set is subsequently fed into the ML classifier.

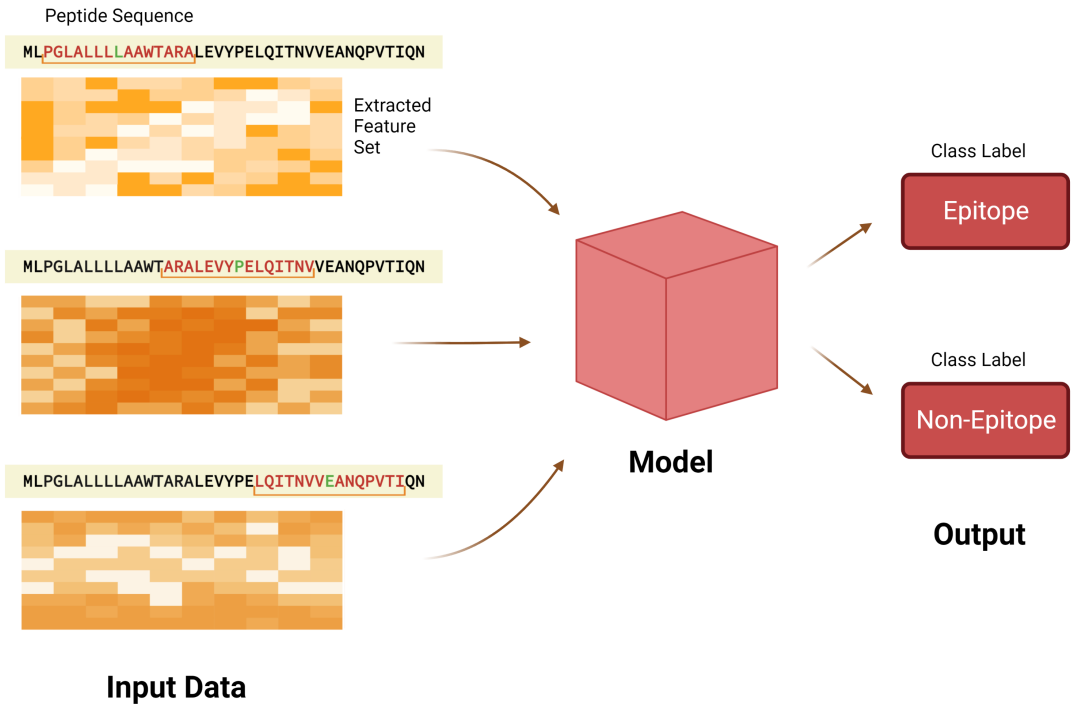


Figure 5. Classification task example. ML models are deployed on the feature spaces associated with each window (input data) to make class predictions. Created in BioRender, March 2023 [7].

Figure 5 shows an outline of the peptide classification task. Here the input data is the feature set extracted from a windowed peptide sequence of the peptide to predict, centred around the amino acid residue currently being classified. The colour matrices of the input data represent the feature space associated with each window (highlighted on the string). New unseen inputs are passed to a pre-trained classification model, which then outputs a predicted label, either 'epitope'or 'non-epitope'. Several classification models may be trained for these types of linear B-cell epitope prediction tasks.

3.3 Exploring Machine Learning Models for Epitope Prediction

Machine learning algorithms build a mathematical model from the data that they are given and this model can then be used to make predictions on new data. There are numerous options for machine learning algorithms to use for epitope prediction. This section explores several modelling approaches for epitope prediction.

3.3.1 Random Forests

Random forests, as introduced by Leo Breiman in 1994, are a form of ensemble learning methods used in machine learning [128, 133]. They function by constructing multiple 'weak'decision tree models and then aggregating their individual predictions to arrive at a final modal class prediction [94, Chapter 12]. This ensemble learning technique is known as bootstrap aggregating, or bagging for short. Random forests offer several advantages over single decision tree models. They are less prone to overfitting, a common issue in machine learning where a model becomes overly specialised to the training data, resulting in poor generalisation on unseen data. Random forests also tend to outperform individual decision trees in classification tasks [94, Chapter 12]. However, random forests are more challenging to interpret compared to single decision tree models due to their relative complexity.

Random forests have been used as classification models in multiple epitope prediction pipelines. Saravanan and Gautham [61] developed an amino acid composition-based feature descriptor: Dipeptide Deviation from Expected Mean (DDE) to distinguish linear B-cell epitopes from non-epitopes. In this study, they evaluated the performance of DDE on an epitope - non-epitope dataset using two machine learning models: a Support Vector Machine and an AdaBoost-Random Forest. Using 5-fold cross-validation the overall reported accuracy of the support vector machine model using DDE was 65% with an F1 score of 0.64 and MCC of 0.213 and the overall accuracy of the AdaBoost random forest model was 69.12% with an F1 score of 0.69 and an MCC of 0.386 [61]. Another

notable application of random forests in epitope prediction comes from Jespersen, Peters, Nielsen, and Marcatili [63]. Their web server, B-cell epitope prediction (BepiPred 2.0) is based on a random forest algorithm and has been shown to outperform other epitope prediction pipelines (including LBtope, BCPREDS and CBTOPE) [63]. This demonstrates the effectiveness of random forests as tools for epitope prediction, contributing to improved accuracy and reliability in this critical field of research.

3.3.2 Support Vector Machines

Support Vector Machines (SVMs) [127, 131] are versatile supervised learning models, that can handle both classification and regression challenges. SVMs operate by mapping input data into a high-dimensional space and then identifying a hyper-plane that maximizes the margin between the two classes [137]. Renowned for their efficiency in high-dimensional feature spaces, SVMs are a favored machine learning algorithm for classification tasks, thanks to their potent performance and reasonable computational demands. Consequently, they frequently find application in machine learning-based epitope prediction pipelines, as demonstrated in a variety of studies [47, 48, 51, 52, 54, 56, 58, 62] (Table 1).

Support Vector Machines have been used in numerous epitope prediction pipelines including: Chen et al. (2007) [47] who introduced an innovative method for predicting linear B-cell epitopes, leveraging an amino acid pair antigenicity scale and SVMs. This method predicts 20-mer peptides and reports, using 5-fold cross-validation, maximum prediction accuracies ranging from 64.39 % to 68.07 %. A later study by EL-Manzalawy, Dobbs and Honavar (2008) [48] focused on predicting linear B-cell epitopes directly from sequence data using a string kernel-based SVM. This study "evaluated support vector machine classifiers trained utilising five different kernel methods using 5-fold cross-validation" on a linear B-cell epitope dataset. Based on their findings they proposed a method for predicting linear epitopes, BCPred, which achieved an estimated prediction accuracy of 67.9 % and area under the receiver operating characteristic (ROC) curve of 0.758. These investigations highlight the capabilities of SVMs in the realm of epitope prediction.

3.3.3 Artificial Neural Networks

Artificial neural networks (ANNs) [129, 130], are ML models inspired by biological neural networks. Examples of types of feed-forward ANNs include the Multilayer Perceptron (MLP). MLP's are characterised by a specific architectural layout. Generally, their structure comprises an initial input layer that receives the input data, followed by one or more hidden layers and finally an output layer that generates the classification outcomes or predictions. MLP's have fully connected layers: every node within the network is connected to every

node in the layer that precedes it and comes after it. The input layer passes the input vectors to the neural network, the hidden layer(s) are where the transformations are performed and the output layer makes a prediction based on the input [138]. MLPs are renowned for having strong generalisation capabilities, allowing them to effectively model complex, non-linear functions. Unlike many other algorithms, MLPs make no underlying assumptions regarding the distribution of the data they operate on, providing a versatile and adaptable tool for diverse machine learning tasks [139]. Neural networks have also been used in several epitope prediction works [38, 60, 65, 67].

3.3.4 Gradient Boosting / XGBoost

Boosting is a machine learning ensemble method [140] that combines the predictions of multiple 'weak'individual models to create a more accurate final prediction. Unlike methods like bagging, where individual models are trained independently, boosting aims to improve model performance by taking into account the performance of previous models [94, Chapter 12.4]. This iterative approach helps the ensemble focus on the samples that are more challenging to classify correctly. Gradient boosting is a specific type of boosting technique that builds multiple models sequentially, with each new model attempting to correct the errors made by the previous ones, these models are often decision trees.

XGBoost (Extreme Gradient Boosting) [141] is an implementation of the gradient boosting framework, specifically designed for solving classification, regression, and ranking problems. This powerful algorithm is known for its efficiency and effectiveness in producing highly accurate predictions. Tree boosting, the technique employed by XGBoost, has gained significant popularity in the machine learning community due to its ability to deliver state-of-the-art results on various classification benchmarks [141, 142]. XGBoost, in particular, has a remarkable track record of success in numerous machine learning challenges and competitions. Gradient boosting has been used for epitope prediction, Manavalan et al., (2018) used gradient boosting in their work *"iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction"* [39]. In their research, they employed gradient boosting as part of an ensemble model for epitope prediction, which also included Extremely Randomised Trees (ERT). The reported predictive performance of their ensemble model was a Matthews correlation coefficient (MCC) of 0.463 when evaluated through cross-validation on an independent test set. This performance showcases the potential effectiveness of gradient boosting in enhancing linear B-cell epitope prediction.

3.4 Evaluating Classifier Performance - Performance Measures

Evaluating the performance of a classification model is essential to understand its potential for generalising to previously unseen data. Performance evaluation of a classifier is conventionally conducted using a separate dataset known as the test set. The test set comprises records that the model has not been exposed to during the training phase: the test set data is distinct from the training dataset. Prior to modelling experimentation, it is essential to set aside the test set. When curating a dataset for a machine learning problem, it is standard practice to partition the labelled data into two distinct subsets: the training set (for classifier training), and the test set (for model evaluation). To obtain valuable insights into the generalisation performance of the trained classifiers, it is vital for the test set to be representative of both the training data and any unseen real-world data that the classifier is intended to make predictions on. Having an independent and representative test set ensures that the evaluation results accurately reflect the classifier's ability to perform well on new and unseen data, allowing for a reliable assessment of its effectiveness in real-world scenarios.

After training, a classification model can be used to make predictions on unseen test instances. The classifier assigns a class label to each test instance based on the patterns and features it has learned during training. These test set predictions, output by the classifier, are then compared with the actual class labels (ground truth) to estimate the classifier's generalisation performance (ability to correctly predict linear B-cell epitopes). To better visualise the performance of a classification model in classifying test instances, a confusion matrix is often employed.

	Predicted Class		
Actual Class		**Positive**	**Negative**
	Positive	True Positive (TP)	False Negative (FN)
	Negative	False Positive (FP)	True Negative (TN)

Table 4. Confusion Matrix

Table 4 shows the structure of a standard confusion matrix. Confusion matrices display a breakdown of how the predictions made by a classification model compare to the ground truth/actual class labels of the test instances. As displayed in Table 4, a confusion matrix is typically organised into rows and columns where each row corresponds to the actual class labels and each column corresponds to the predicted class label. The counts of instances in each combination of true and predicted labels are recorded in the cells of the matrix. Considering the confusion matrix in Table 4 the four cells contain the counts of:

- **True Positive (TP):** Correctly predicted positive.
- **False Negative (FN):** Instances predicted to be negative that are actually positive.
- **False Positive (FP):** Instances predicted to be positive that are actually negative.
- **True Negative (TN):** Correctly predicted negative.

Confusion matrices are useful for understanding the types of errors made by classifiers and for evaluating their performance across different classes. Analysing classification results in a confusion matrix can help to identify specific areas where the classifier is miss-classifying labels more frequently, this can guide model refinement or dataset balancing techniques, to try and achieve better classification performance. Confusion matrices can also be used to derive other performance indicators, like accuracy and sensitivity, using the values in the matrix.

3.4.1 Defining Selected Performance Measures

The performance measures selected to evaluate classification models in this work are defined below:

- Accuracy:
$$\frac{TP + TN}{TP + FP + TN + FN}$$

- Sensitivity:
$$\frac{TP}{TP + FN}$$

- Specificity:
$$\frac{TN}{TN + FP}$$

- Positive Predictive Value:
$$\frac{TP}{TP + FP}$$

- Negative Predictive Value:
$$\frac{TN}{TN + FN}$$

- F1 Score:
$$2 \cdot \frac{PPV \cdot TPR}{PPV + TPR}$$

- MCC:
$$\frac{(TP \cdot TN) - (FP \cdot FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

Accuracy (ACC): Calculates the percentage of correctly classified labels made by the classifier. Accuracy is a basic and easy to interpret measure, though it is sensitive to class balance and therefore can be misleading if reported without considering the class balance.

Sensitivity (SENS): Also known as the True Positive Rate (TPR) is the ability of a classifier to correctly identify (or 'recall') positive records. Sensitivity is therefore a valuable measure in applications like medical diagnosis where identifying positive records is critical. A high sensitivity indicates the classifier's ability to correctly identifying positive cases. However, achieving high sensitivity might lead to an increase in false positives, which is an important trade-off to consider.

Specificity (SPEC): Also known as the True Negative Rate (TNR) is the ability of a classifier to correctly identify negative instances. It is particularly valuable when minimising false positive predictions is a priority.

Positive Predictive Value (PPV): Also known as Precision, is the proportion of positive predictions that are true positive predictions. It is a measure of the ability of a classifier to avoid false positive predictions. PPV is particularly useful when false positives are undesirable, however it's important to consider that this measure does not take into account false negative predictions.

Negative Predictive Value (NPV): The proportion of negative predictions that are true negative predictions.

F1 Score (F1): The harmonic mean of precision (PPV) and recall (SENS), useful for balancing precision (PPV) and recall (SENS) as it gives equal weight to both. This makes the F1 score a good measure for comparing the performance of different classifiers as it gives a good idea of the performance in terms of both precision and recall. However, the F1 score can be sensitive to class imbalance.

Matthews Correlation Coefficient MCC: A measure of the quality of binary classifications. It's a special case of the ϕ phi coefficient [143] developed by Matthews in 1975 for comparison of chemical structures [144]. An MCC score is a value between -1 and 1, with 1 representing a perfect prediction, 0 no better than random and -1 total disagreement between prediction and observation. MCC is not sensitive to class imbalance, it is a balanced measure which can be used even if classes are of very different sizes.

Area Under the ROC Curve (AUC): The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) is another measure used to assess the quality of binary predictions. This measures the model's ability to distinguish between the positive and negative classes. It is calculated by first plotting an ROC curve (plotting sensitivity against 1-specificity) at various threshold values, and then calculating the total area under the ROC curve. An AUC score is a value between 0 and 1, with 1 representing a perfect prediction, 0.5 no better than random and 0 total disagreement.

Selecting appropriate performance measures for this machine learning task is crucial because it directly impacts how the model's performance and suitability for epitope prediction is evaluated. This work will employ the performance measures outlined above (ACC, SENS, SPEC, PPV, NPV, F1, MCC & AUC) when assessing classifiers for linear B-cell epitope prediction. Beyond their described merits, these performance indicators also facilitate comparisons with other predictors reported in the existing literature.

4. Organism-Specific Modelling for Linear B-Cell Epitope Prediction

Most current epitope prediction tools were developed under a generalist approach, training models using heterogeneous datasets to create predictors that can be deployed for a wide variety of pathogens. However, with the rapid advancements in processing power and the ever-increasing availability of epitope data spanning a diverse range of pathogens, there arises an opportunity to explore organism or taxon-specific models as a viable alternative. This approach holds untapped potential to significantly improve predictive performance.

This chapter explores organism-specific training for epitope prediction models and its potential to yield substantial performance gains. Organism-specific models are compared to models trained with heterogeneous and hybrid data, as well as widely-used predictors from existing literature, across various performance indicators, to discern the advantages of organism-specific training. The findings of this research showcase a promising alternative for developing custom-tailored predictive models with exceptional predictive power. Moreover, these models offer ease of implementation and deployment, specifically tailored to investigate pathogens of interest.

4.1 Generalist Modelling Approaches for Epitope Prediction

Existing epitope prediction tools, to the best of our knowledge, rely on datasets comprising labeled peptide sequences derived from a diverse range of organisms (Table 5). The use of such heterogeneous datasets stems from the common objective of developing general-purpose predictors that can be readily employed without requiring users to specify the source organism of the submitted peptides for classification. In fact, as recently as 2020, Collatz et al. [65] argued that including "*a large variety of known epitopes from evolutionarily distinct organisms in the training set*" is crucial for achieving unbiased classification. Even some of the most recently published epitope predictors, such as EpitopeVec [66] and BepiPred 3.0 [67], continue to utilise phylogenetically heterogeneous training sets. This assumption holds merit when aiming to build generalist, one-size-fits-all models that cater to a wide array of scenarios. However, it may prove unnecessary or even counterproductive if the model is intended to generalise solely to a specific subset of all possible observations.

Table 5. Epitope prediction tools in the literature and the composition of their training datasets. PK = prokaryotes; VI = virus; FG = fungi; PR = protozoan; HM = human; OE = other eukaryotes. Details on the source databases used can be found on the works cited in the *Method* column.

Method	**Year**	**Training dataset**	
		Sources	**Composition**
Antigenic [36]	1990	-	-
PREDITOP [43]	1993	-	-
People [44]	1999	-	-
Bepitope [45]	2003	-	-
Bcepred [46]	2004	-	-
Söllner/Mayer [37]	2006	BCIPEP, FIMM	PK+VI+FG+PR+OE
ABCpred [38]	2006	BCIPEP	PK+VI+FG+PR+OE
Chen [47]	2007	BCIPEP	PK+VI+FG+PR+OE
BCpred [48]	2008	BCIPEP	PK+VI+FG+PR+OE
FBCpred [48]	2008	BCIPEP	PK+VI+FG+PR+OE
LEPD [49]	2008	Antijen, Pellequer	PK+VI+PR+HM+OE
Epitopia [50]	2009	BCIPEP	PK+VI+FG+PR+OE
COBEpro [51]	2009	BCIPEP, HIV	PK+VI+FG+PR+OE
BayesB [52]	2010	BCIPEP	PK+VI+FG+PR+OE
LEPS [53]	2011	BCIPEP	PK+VI+FG+PR+OE
BEOracle [54]	2011	BCIPEP, IEDB, Antijen	PK+VI+FG+HM+OE
SVMTriP [132]	2012	IEDB	PK+VI+PR+HM+OE
BEST [56]	2012	BCIPEP, SWISS-PROT	PK+VI+OE
LBtope [57]	2013	IEDB	PK+VI+HM+PR+OE
BeePro [58]	2013	Mix of datasets	PK+VI+FG+PR+HM+OE
EPMLR [59]	2014	BCIPEP, IEDB, Antijen	PK+VI+FG+HM+OE
DMN-LBE [60]	2015	IEDB	PK+VI+HM+PR+OE
LBEEP [61]	2015	IEDB	PK+VI+HM+PR+OE
APCpred [62]	2015	BCIPEP	PK+VI+FG+PR+OE
BepiPred 2.0 [63]	2017	PDB	PK+VI+OE
DRREP [145]	2017	BCIPEP	PK+VI+FG+PR+OE
iBCE-EL [146]	2018	IEDB	PK+VI+HM+PR+OE
EpiDope [65]	2020	IEDB	PK+VI+HM+PR+OE
EpitopeVec [66]	2021	IEDB, Bcipep	PK+VI+HM+PR+OE
BepiPred 3.0 [67]	2022	PDB	PK+VI+OE

4.2 The Organism Specific Hypothesis

The continuous advancements in computational processing power, coupled with the increasing abundance of data available for distinct pathogens, suggest that adopting organism or taxon-specific models may become a feasible alternative for linear B-cell epitope prediction. Rather than relying on a single generalist model, developing predictors specifically tailored for individual pathogens holds potential advantages. Developing organism-specific predictors would involve training models using smaller, yet potentially higher-quality, datasets, resulting in improved predictive performance for new epitopes related to the target organism, and possibly even its phylogenetically close relatives. Under this alternative approach of training bespoke models for distinct pathogens (or groups of pathogens), the objective is to obtain predictors that generalise well to the target organism(s), rather than to the whole variety of pathogens that may interact with a given host.

This study investigates the effects of employing organism-specific datasets to train machine learning models for linear B-cell epitope prediction. Proof-of-concept predictors are trained using organism-specific, heterogeneous, and hybrid data, relating to data-rich pathogens representing two major classes of parasitic organisms: nematodes and viruses. The effects of these training sets on the models' generalisation performance is quantitatively assessed to ascertain whether organism-specific training can indeed yield superior predictors. The results obtained from three test cases not only support the viability of this approach, but also demonstrate that even relatively simple models trained on organism-specific data exhibit superior performance compared to current state-of-the-art predictors, as evaluated across multiple performance measures.

4.3 Organism Specific Modelling Pipeline Overview

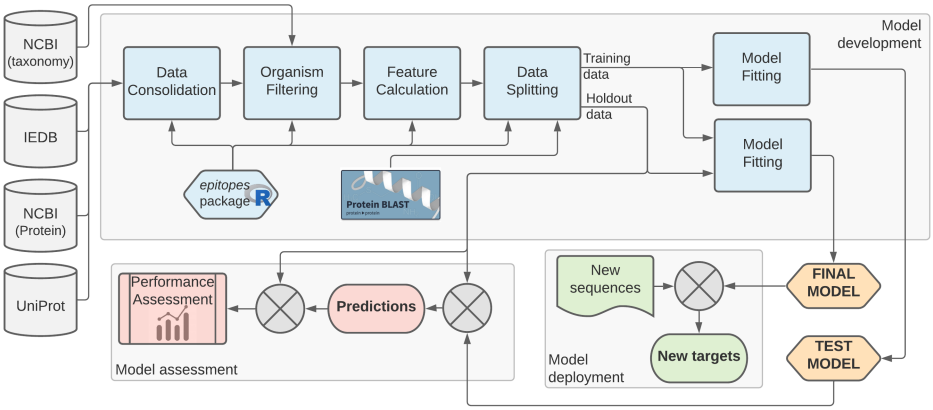


Figure 6. Overview of the Epitope Prediction Pipeline.

Figure 6 illustrates the organism-specific epitope prediction pipeline used in this work. To construct an organism-specific dataset, publicly-available data is retrieved from the IEDB [28], NCBI [109] and UniProt [22]. All records relating to known epitopes and known non-immunogenic peptides from the target organism(s) are retrieved from the IEDB and additional protein information is retrieved from NCBI and UniProt. The NCBI Taxonomy database was also used to retrieve taxonomic/phylogenetic information relating to the entries. 845 simple features (described in section 2.6.2) are then calculated for each amino acid in a record in the dataset. These features are based on the local neighbourhood of every position extracted using a 15-AA sliding window representation with a step size of one (Section 2.6.1, Figure 4). Subsequently, the data is divided at the protein level, using protein ID and similarity, into a training set for model development and a hold-out set for estimating the models' generalisation performance (detailed in Section 4.4.2). The *epitopes* R package, which implements the main data retrieval and consolidation elements of this pipeline, can be accessed at https://github.com/JodieAsh/Epitopes.git.

4.4 Organism Specific Dataset Generation

4.4.1 Target Pathogens

The selection of pathogens for this study was primarily based on the availability of a substantial volume of validated positive and negative observations within the IEDB. This criterion facilitated the implementation of the rigorous validation strategy outlined earlier, which involved the use of a 25% hold-out set (see Section 4.4.2) while still ensuring sufficient data for model development. To identify pathogens meeting these requirements, the ten organism IDs with the greatest number of valid entries in the IEDB were extracted, following the filtering process described in Section 2.5. The selection process involved a further two key considerations:

(i) Ensuring a reasonable balance between positive and negative examples, thereby excluding entries with heavily imbalanced class distributions (IDs: 353153, 1314, 5833, 1392).
(ii) Focusing on representing pathogens of interest. This removed entries related to allergens or potential self-epitopes (IDs: 9606, 9913 and 3818).

Based on these criteria three pathogens of interest were selected: a multi-cellular parasite (*Onchocerca volvulus*), an RNA virus (Hepatitis C Virus) and a DNA virus (Epstein-Barr Virus). This selection allowed for the evaluation of the pathogen-specific tools developed across distinct classes of organisms.

A brief summary of each of the selected target organisms (used to investigate the efficacy of organism-specific training) is provided below:

- *Onchocerca volvulus* (taxonomy ID: 6282): This roundworm (Nematoda) is the causative agent of Onchocerciasis, a major cause of blindness worldwide [147]. The disease affects over 37 million people, primarily in Africa and Latin America [148, 149].

- Epstein-Barr Virus (taxonomy ID: 10376): This double-stranded DNA virus belongs to the *Herpesviridae* family. It is the causative agent of infectious mononucleosis and has been linked to various human neoplastic diseases [150].

- Hepatitis C Virus (taxonomy ID: 11102): This positive-sense single-stranded RNA virus belongs to the *Flaviviridae* family. It is responsible for causing Hepatitis C and has been linked to the development of certain cancers [151].

4.4.2 Dataset Generation

Organism specific datasets were generated for each selected pathogen (*Onchocerca volvulus*, Epstein-Barr Virus and Hepatitis C Virus) based on the full XML export of the Immune Epitope DataBase (IEDB) [28] retrieved on 10 October 2020, and filtered according to the criteria outlined in Section 2.5.

For each pathogen, several distinct datasets were generated as follows:

1. Extraction of specific pathogen data: all examples related to the specific pathogen were extracted based on the taxonomy ID information from the IEDB data. This included all taxonomically-dependent IDs, such as those related to subspecies or strains, as part of the data.

2. **Hold-out dataset**: To ensure unbiased model evaluation, a subset of the organism-specific data (comprising approximately 25 % of the available observations) was set aside as a (*Hold-out*) set. This test data was not utilised at any point during the model development process, including during exploratory data analysis, data pre-processing, modelling and hyperparameter tuning.

3. **Organism-specific training dataset**: The remaining subset, which contained 75% of the labelled peptides specific to the target pathogen, was designated as the organism-specific (*OrgSpec*) training set. To minimise the chances of data leakage [152] the division of datasets was conducted at the protein level. This was based on protein ID, as well as sequence coverage and similarity. Proteins with similarity and/or coverage greater than 80% were always placed within in the same split, therefore preventing data leakage due to both peptide and protein similarities.

4. **Heterogeneous training dataset**: A second training dataset (*Heterogeneous*) was created by randomly sampling observations (grouped by taxonomy ID) from the complete IEDB database. Notably, this sampling excluded any observations associated to the target pathogen. To ensure a balanced training set, the sampling routine included as many organisms as necessary to construct a class-balanced *heterogeneous* training set, containing between 2,000 and 3,000 labelled peptides of each class (epitope/non-epitope).

5. **Hybrid training dataset**: Finally, a third training set (*Hybrid*) was assembled by combining the *OrgSpec* and *Heterogeneous* sets, incorporating both organism-specific data and a diverse range of heterogeneous peptides from other organisms.

The dataset sizes extracted from each organism are documented in Table 6.

		OrgSpec	Heter.	Hybrid	Hold-out
O. volvulus	Positive	2441	2634	5075	832
	Negative	2378	2922	5300	777
Epstein-Barr	Positive	1746	1981	3727	625
	Negative	811	1864	2675	315
Hep. C Virus	Positive	919	1926	2845	218
	Negative	783	1975	2758	358

Table 6. Number of positive/negative examples in each organism-specific dataset. Hold-out sets always contain only target organism proteins that were not seen during model training.

In each scenario, the *Hybrid* dataset was the largest one, followed by the *Heterogeneous* dataset and then the *Organism-Specific (OrgSpec)* dataset. This arrangement was intentionally structured to investigate the hypothesis that prioritising sample relevance (represented by data from the organism of interest for which the models are being developed) over sample size (which increases when heterogeneous observations are included) would lead to improved prediction performance.

4.5 Outline of Main Investigations

The datasets assembled for each pathogen facilitated the following investigations:

(i) The examination of the generalisation performance of the models in predicting new epitopes within proteins belonging to the specific organisms for which they were trained. This analysis involved assessing the predictive performance on the distinct proteins that were reserved as the organism-specific *Hold-out* set.

(ii) The investigation of the effect of using exclusively organism-specific data on predictive performance. This was accomplished by comparing models developed using the *OrgSpec*, *Hybrid* and *Heterogeneous* datasets. Given that all pre-processing, feature development, and classification models were consistent across cases, any systematic differences in performance could be attributed to the pre-selection of training data.

(iii) The comparison of the performance of organism-specific models against conventional approaches found in the literature. The inclusion of a hold-out approach (over cross-validation) for model assessment was particularly important for this purpose, as it allowed for the estimation of the generalisation performance of all predictors using the same data. This approach avoided reliance on reported performance values from the literature, which were obtained on distinct datasets or using different testing protocols.

4.6 Modelling

4.6.1 Model Selection

To evaluate the performance of various classification models for organism-specific epitope prediction, set in the space of features defined in this work, the *O. volvulus* (organism-specific) training dataset was used. The dataset was further split into training and validation sets based on protein ID and sequence similarity, excluding the hold-out set used in the main experiments. The following models were tested: Random Forest (RF) [133], XGBoost [141], Support Vector Machine (SVM) [131] and Multilayer Perceptron Neural Network (MLP) [153]. These models were assessed across multiple performance measures, including Accuracy, Matthews Correlation Coefficient (MCC), Area Under the ROC curve (AUC) and Positive Predictive Value (PPV).

For these model selection experiments, all classification models were implemented using the Scikit-learn package version $0.24.1$ [154], except for XGBoost, which utilised the implementation available in the XGBoost package [155]. The default hyperparameter values provided by these packages were initially used for all methods. The specific hyperparameters for each model can be found in Table 7. The full Python modelling pipeline is available at: https://github.com/JodieAsh/Epitope_Prediction_V2.git.

Table 7. Model hyperparameters used in the model selection experiments. Default values from each implementation were used.

Model	**Parameter: value used**
Random Forest	*n_estimators*: 100
	split criterion: Gini
	max_depth: None
	min_samples: 2
	min_samples_leaf: 1
	min_weight_fraction_leaf: 0
	max_features: auto
	max_leaf_nodes: None
	min_impurity_decrease: 0
	min_impurity_split: None
	bootstrap: True
	warm_start: False
	class_weight: None
	ccp_alpha: 0
	max_samples: None
XGBoost	*eta*: 0.3
	gamma: 0
	max_depth: 6
	min_child_weight: 1
	max_delta_step: 0
	subsample: 1
	sampling_method: auto
	colsample_bytree: 1
	lambda: 1
	alpha: 0
	tree_method: auto
	sketch_eps: 0.03
	scale_pos_weight: 1
	updater: grow_colmaker
	refresh_leaf: 1
	grow_policy: depthwise
	max_leaves: 0
	max_bin: 256
	predictor: auto
Support Vector Machine	*C*: 1

Table 7 – continued from previous page

Model	Parameters: values used (package defaults)
	kernel: rbf
	degree: 3
	gamma: scale
	coef0: 0
	shrinking: True
	probability: True
	tol: 1e-3
	cache_size: 200
	class_weight: None
	decision_function_shape: ovr
	break_ties: False
Multilayer Perceptron	*hidden_layers_sizes*: (100,)
	activation: relu
	solver: adam
	alpha: 0.0001
	batch_size: auto
	learning_rate: constant
	learning_rate_init: 0.001
	power_t: 0.5
	max_iter: 200
	shuffle: True
	tol: 1e-4
	warm_start: False
	momentum: 0.9
	nesterovs_momentum: True
	early_stopping: False
	validation_fraction: 0.1
	beta_1: 0.9
	beta_2: 0.999
	epsilon: 1e-8,
	n_iter_no_change: 10
	max_fun: 15000

After evaluating the performance of each model across multiple indicators, the Random Forest model was found to consistently outperformed the other models. The results, presented in Table 8, show that the Random Forest model achieved the highest scores across all performance measures. As a result, the Random Forest model was selected as the primary modeling approach for all experiments.

Table 8. Point estimates of performance obtained for initial model exploration (on the organism-specific validation set). ACC = Accuracy, MCC = Matthews Correlation Coefficient, PPV = Positive Predictive Value, AUC = Area Under the ROC Curve.

Model	**ACC**	**MCC**	**PPV**	**AUC**
Random Forest	0.695	0.377	0.739	0.753
XGBoost	0.690	0.364	0.727	0.741
Support Vector Machine	0.674	0.334	0.719	0.722
Multilayer Perceptron	0.661	0.300	0.694	0.705

Based on the results obtained, which demonstrated the superior performance of the Random Forest model across various performance measures, it was decided to utilise the Random Forest algorithm for developing all organism-specific models in this work. While additional comprehensive model investigations could be conducted, the achieved high performance was deemed satisfactory, justifying the adoption of Random Forest as the model of choice for this study.

Random Forests, are ensemble learning methods that harness the strength of multiple weaker decision tree (DT) models to produce a collective output based on the combined predictions of the underlying DTs. Random forests offer a good balance between computational cost/efficiency and performance; they are robust and versatile across diverse data types and scales, making them an ideal choice for applications like epitope prediction [61, 63]. Table 8 showed that preliminary comparative testing favored Random Forests and Gradient Boosting models over alternatives such as multi-layer perceptron neural networks and SVMs. While Gradient Boosting exhibited competitive performance, the computational demands associated with it were comparatively higher. In contrast, Random Forests demonstrated a favorable trade-off between computational efficiency and predictive accuracy. This distinction in computational costs solidified RF's position as the preferred choice, ensuring a feasible and efficient modeling process.

4.6.2 Hyperparameter Tuning

To potentially enhance the performance of the Random Forest classification model, hyperparameter tuning was investigated. Similar to the model selection experiments, the tuning experiments were conducted using the *O. volvulus* training dataset (further split into training and validation). These tuning tests were performed using Scikit-learn's randomised search function, using the following fixed parameters:

- *n_iter*: 200
- *scoring*: mcc_scorer
- *refit*: True
- *error_score*: np.nan
- *pre_dispatch*: None

The following hyperparameters were tuned for the Random Forest model:

- *bootstrap* $\in \{False, True\}$. **Selected value:** True
- *max_features* $\in \{auto, sqrt\}$. **Selected value:** auto
- *min_samples_split* $\in \{10, 15, 20, 25, 30\}$. **Selected value:** 20
- *n_estimators* $\in \{500, 550, 600, \ldots, 1000\}$. **Selected value:** 650
- *min_samples_leaf* $\in \{2, 3, 4, 5, 6\}$. **Selected value:** 4
- *max_depth* $\in \{20, 40, \ldots, 120\}$. **Selected value:** 80

After tuning these hyperparameters, the final performance values are presented in Table 9. For convenience, the baseline results obtained by the standard configuration are also included in this table.

Table 9. Random Forest performance indicators before and after parameter tuning.

Method	**ACC**	**MCC**	**PPV**	**AUC**
Benchmark	0.695	0.377	0.739	0.753
After Tuning	0.703	0.394	0.744	0.760

By exploring various combinations of hyperparameters using this approach, the aim was to identify the optimal configuration that maximises the performance of the Random Forest model. It cannot be definitively concluded that hyperparameter tuning does not significantly impact the performance of the method. Nonetheless, based on the observations, it was noted that, even when employing considerably different hyperparameter values, these variations did not result in substantial differences compared to the default. Considering this, a decision was made to exclude hyperparameter tuning from the final implementation of the models in this work. This choice was based on adopting a lower-complexity approach to pipeline design and the rationale that the primary objective of the research was to showcase the organism-specific training principle rather than extensively explore model

optimisation. However, it's important to note that this outcome does not preclude future, more comprehensive investigations of different models and hyperparameter settings, which could be valuable for fine-tuning performance in specific contexts.

4.6.3 Dimensionality Reduction

For the organism specific pipeline, various dimensionality reduction techniques were investigated, including:

(i) Filter methods based on extracting the top *K* features using Mutual Information [106, 156] and the Anova F-value [157] as ordering scores.
(ii) Principal Component Analysis (PCA) [158, 159].
(iii) A wrapper method, Maximum Relevance-Minimum Redundancy (MRMR) [107].

These dimensionality reduction methods were assessed using the *O. volvulus* dataset. As mentioned, to ensure unbiased evaluation, the data was split into training and validation sets based on protein ID and sequence similarity. The holdout set, used in the main experiments, was excluded from these evaluations. In this investigation, all dimensionality reduction techniques were implemented using Scikit-learn. The default parameters provided by Scikit-learn for each technique were employed during evaluation, as listed in Table 10. The primary objective of this investigation was to identify the most effective dimensionality reduction technique that could enhance epitope prediction performance. By comparing the results obtained from each method, the study aimed to determine which technique provided the most significant improvement in epitope prediction performance.

Table 10. Default Dimensionality Reduction Techniques Parameters

Method	**Parameters**
Principal Component Analysis	*n_components*$\in \{0.95, 0.5, 0.15\}$ *copy*: True *whiten*: False *sdv_solver*: auto *tol*: 0 *iterated_power*: auto
Mutual Information	*discrete_features*: auto *n_neighbors*: 3 *copy*: True
Select K Best	*score_func*: f_classif *k*: 15
MRMR	*K*: 15 *relevance*: f *redundancy*: c *denominator*: mean *only_same_domain*: False

The evaluation of various dimensionality reduction techniques did not yield significant performance gains for any of the classifiers tested. As an illustrative example, Table 11 displays the results obtained specifically for the default Random Forest classifier.

Method	**Features**	**ACC**	**MCC**	**PPV**	**AUC**
Benchmark	845	0.695	0.377	0.739	0.753
Mutual Information	15	0.688	0.365	0.710	0.746
PCA (0.95)	521	0.655	0.284	0.683	0.699
PCA (0.50)	121	0.674	0.334	0.718	0.726
PCA (0.15)	11	0.678	0.353	0.740	0.734
Select K Best	70	0.691	0.370	0.736	0.745
MRMR	15	0.700	0.385	0.738	0.747

Table 11. Random Forest classifier performance after dimensionality reduction. The PCA values refer to the proportion of variance retained. The baseline values obtained by the Random Forest on the full feature set are repeated here for convenience (referenced to as *benchmark*).

Despite experimenting with different dimensionality reduction methods, the predictive performance of the classifiers remained relatively stable, and no substantial improvements were observed in the reduced feature sets. The study did not identify any dimensionality reduction technique that significantly enhanced the performance of the classifiers in the context of epitope prediction.

Given that the primary aim of this research was to demonstrate the impact of organism-specific training, rather than building a fully-deployed pipeline, the decision was made to retain the full feature sets in the solutions presented. This choice was based on the observed lack of improvement in generalisation performance with feature reduction (Table 11) and the understanding that Random Forest inherently performs its own embedded feature prioritisation process. Reducing the feature space through dimensionality reduction will, however, prove valuable when developing a user-facing interface for the organism-specific pipeline. It can effectively reduce the computational costs associated with feature calculation and model fitting, making the pipeline more efficient and user-friendly without compromising its performance.

4.7 Model Testing

After experimentation with modelling, hyperparameter tuning and feature selection, all training datasets from Section 4.4.2 were used to develop Random Forest (RF) predictors [133]. Multiple RF models were built using the training datasets (extracted from a specific target pathogen). Once trained, each RF model was deployed to make predictions on the corresponding hold-out set reserved specifically for that particular pathogen. This approach ensured that predictions were made for epitopes in proteins belonging to the respective target organisms for which the models were developed. By evaluating the performance of each RF model on its hold-out set, the predictive capability of the models for their corresponding pathogens could be assessed.

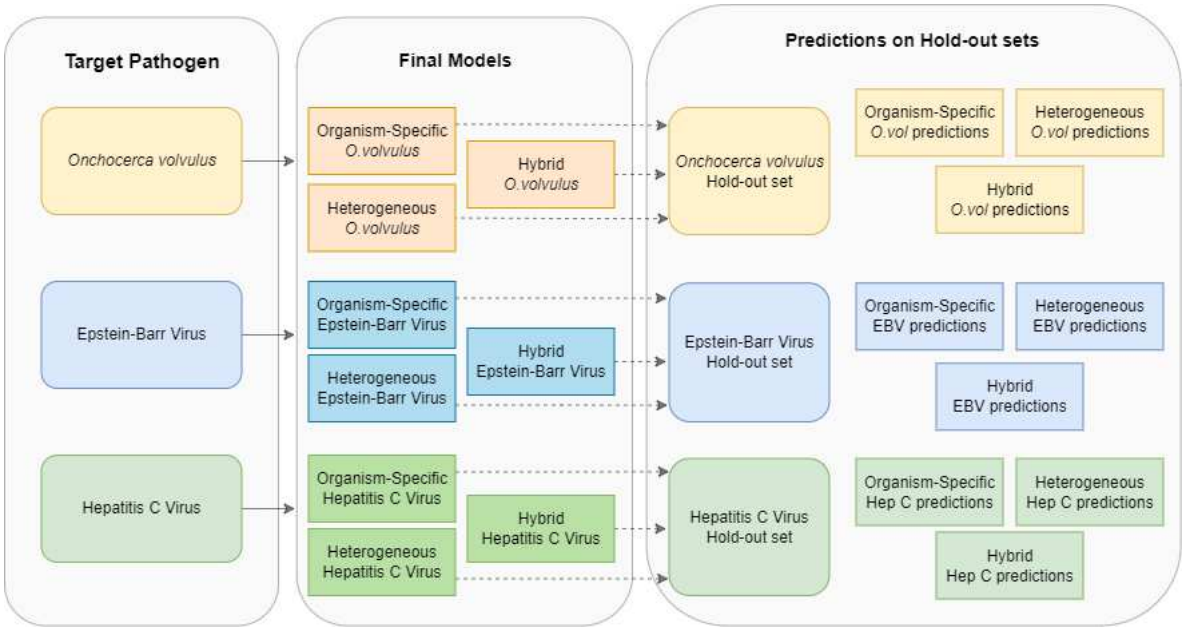


Figure 7. Summary of final models and predictions generated. Three separate models were trained for each target pathogen: an Org-Spec, Heterogeneous and Hybrid model. All were deployed on the pathogen specific hold-out sets.

Figure 7 provides an overview of the final models generated for each of the target pathogens and their predictions. Each model was employed to make predictions on the hold-out set reserved for its corresponding pathogen, resulting in three sets of predictions (one from each model) for each pathogen. The predictions generated by each model comprised a predicted probability for every position within each protein of the hold-out set, indicating whether that position belonged to an epitope or not. These probabilities were then converted into binary predictions (epitope or non-epitope) using a threshold of 0.5. From these amino-acid wise predictions, predicted epitopes of arbitrary lengths were extracted for each protein in the hold-out set. To minimise prediction noise, positive regions shorter than 8-amino-acids long (consistent with the initial filtering strategy used for the training data) were filtered out from the output. This filtering step helped to ensure that the predicted epitopes were more

accurate and biologically meaningful for further analysis and interpretation. The resulting epitope predictions enabled us to evaluate the performance and predictive capabilities of the organism-specific, heterogeneous, and hybrid models on their corresponding hold-out datasets. By testing the models on unseen data from each pathogen, we could gauge their ability to generalise and make accurate predictions, providing valuable insights into the effectiveness of the organism-specific approach for epitope prediction.

4.7.1 Performance Assessment and Comparison

Several performance indicators were calculated to provide comparability with different references in the literature, and to explore distinct aspects of the predictive behaviour of the models. More specifically, model performance was compared and assessed using: *Positive Predictive Value* (PPV), *Negative Predictive Value* (NPV), *Sensitivity* (SENS), *Accuracy* (ACC), *Area Under the ROC Curve* (AUC) and *Matthews Correlation Coefficient* (MCC). The detailed mathematical definition and interpretation of each of these measures is provided in Section 3.4. By employing these performance indicators and comparing them with existing studies in the field, we obtained a holistic understanding of the models' effectiveness in epitope prediction.

The performance assessment and comparison in this section is exclusively based on out-of-sample predictions, specifically the performance observed on the dedicated *hold-out* set for each individual pathogen. This approach ensures the robustness of the evaluation process, as this hold-out data is entirely excluded from the model development phase. Therefore, the reported performance values serve as reliable estimates of the models' generalisation capabilities for these specific organisms.

Performance was calculated based on peptide-wise correct classifications. Following standard practice, a classification was deemed correct if the model predicted the correct class for more than half of the residues within a labeled peptide. Standard errors of estimation for each performance indicator, along with p-values for comparing mean performance between our reference implementation (trained with *OrgSpec*) and all other comparison methods, were computed using the Bootstrap method [160] (specifically, 999 bootstrap re-samples were generated in all cases). The resulting p-values underwent correction for multiple hypothesis testing (MHT) using the Holm correction technique [161]. This correction strategy ensures stringent control of the Family-wise error rate (FWER) for each hypothesis family. Across all comparisons, significance was established at a collective $\alpha^{\star} = 0.05$ threshold (significance level). To establish a comparison baseline, five widely recognised B-cell epitope predictors with user-friendly online interfaces were employed. These predictors are: Bepipred 2.0 [63], SVMtrip [132], LBtope [57], ABCpred [38], and iBCE-EL [146]. These models were employed to predict epitopes within the same *hold-out* sets, utilising the default configurations of their respective online tools.

4.8 Results

4.8.1 Organism-Specific Training Improves Performance of Linear B-Cell Epitope Prediction

As described in Section 4.7, the performance of the organism-specific Random Forest models (*RF-OrgSpec*) was compared with:

i. Random Forest models using the same parameters but trained with heterogeneous and hybrid data, to investigate the effect of the data selection strategy on performance.
ii. A selection of widely-used predictors from the literature, providing a basis for comparison with established methods.

In all cases the performance evaluation was conducted using the *hold-out* set specifically isolated for each pathogen, which was not used at any point during model development. Figure 8 provides a comprehensive summary of the results obtained for the organisms described in Section 4.4.1. Notably, a strong positive effect of training models with organism-specific data was observed across all datasets. For all studied organisms, (EBV, HepC & *O.volvulus*) the *RF-OrgSpec* models consistently achieved the highest scores among all RF predictors, outperforming the *RF-Hybrid* and *RF-Heterogeneous* models, across all performance measures. A clear performance ordering *RF-OrgSpec* $>$ *RF-Hybrid* $>$ *RF-Heter* can be observed across all pathogens, on all performance indices used, showing that *RF-OrgSpec* outperforms *RF-Hybrid*, which in turn surpasses *RF-Heter*. The statistical analysis, with corrected p-values, further supports these observations, indicating that the observed differences are mostly statistically significant at the joint 0.05 significance level.

These findings serve as strong validation for the initial hypothesis that training models with organism-specific data significantly enhances predictive performance. This improvement is evident even when comparing the performance of the *OrgSpec* models with those that include the same organism-specific data alongside examples from other organisms (*RF-Hybrid*). The results emphasise the effectiveness of organism-specific training in epitope prediction, showcasing its superiority in achieving enhanced predictive capabilities when compared to hybrid and heterogeneous models. These observations highlight the potential of leveraging organism-specific training to develop custom-tailored predictive models with superior performance for specific pathogens.

Figure 8. Performance estimates and standard errors of different predictors on the hold-out data of the target organisms. RF-OrgSpec, RF-Hybrid and RF-Heter models were deployed on the hold-out set along with several widely-used epitope predictors from the literature (SVMtrip, LBtope, iBCE-EL, Bepipred2, ABCpred). The values near each estimate are MHT-corrected p-values for the comparison of mean performance against *RF-OrgSpec*. Estimates are colour-coded for the result of significance tests at the $\alpha^* = 0.05$ significance level (green for significantly worse than *RF-OrgSpec*, red for significantly better, blue for non-statistically significant differences). The p-values were truncated at < 0.01 and > 0.9 due to loss in precision of bootstrap estimates at extreme values.

Additional analyses further confirm that the performance gains of organism-specific prediction are observed specifically for the pathogen on which the model is trained, but are not evident when attempting to predict epitopes for other organisms. Figure 9 contrasts the observed performance of the *RF-OrgSpec* models on the hold-out set of their respective organism with that obtained when predicting epitopes for the other pathogens. The results demonstrate that the substantial gains in organism-specific performance (as shown in Figure 8) come with a trade-off: the models' ability to predict epitopes in proteins from other pathogens is reduced. This observation further supports our underlying hypothesis that organism-specific training enables models to capture unique patterns that are specific to the target pathogen.

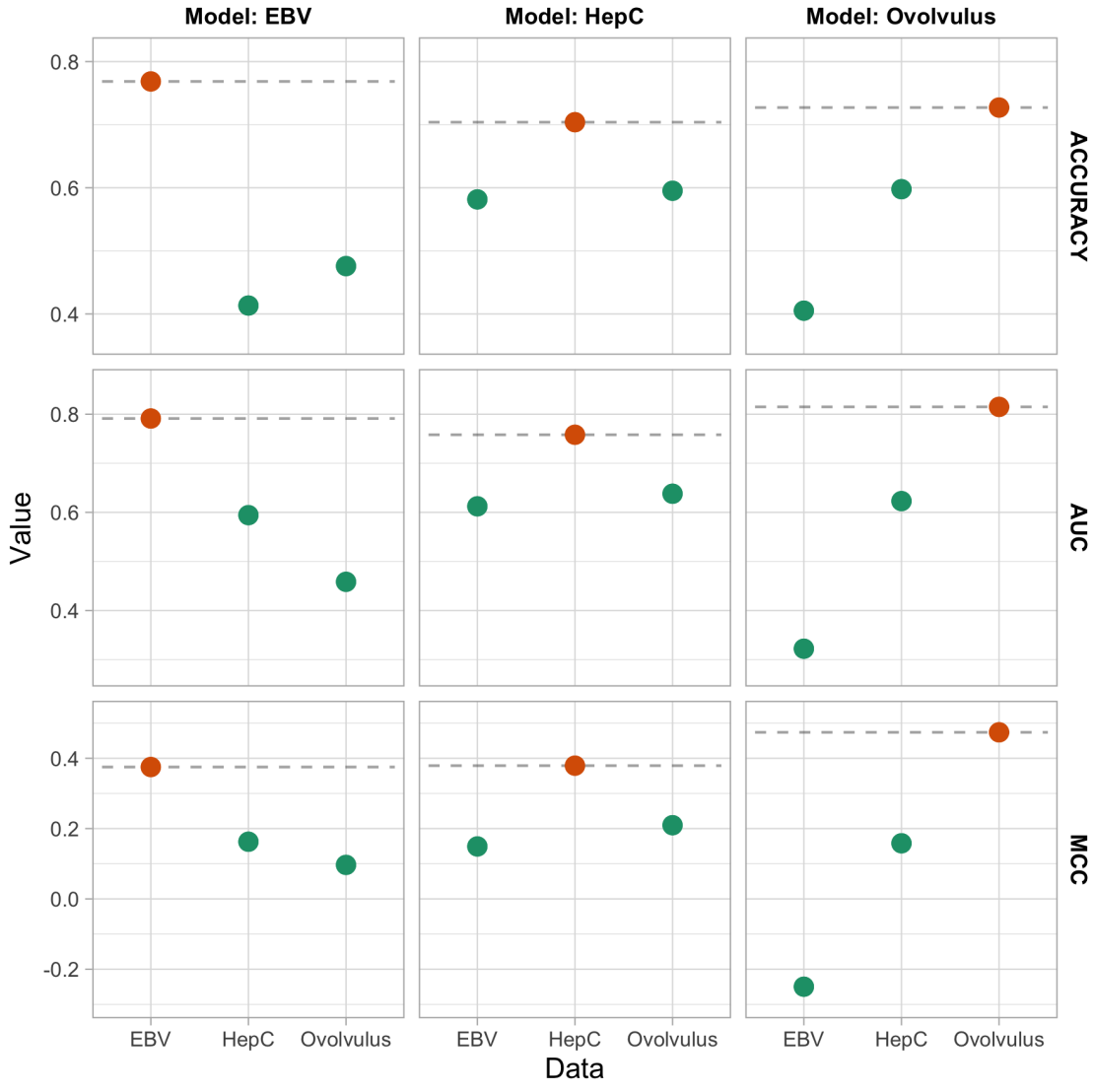


Figure 9. Aggregated performance indices (Accuracy, AUC and MCC) of each organism-specific model on each organism-specific hold-out set. This shows how organism-specific training results in models that generalise very well to the particular pathogen for which they were developed, at the cost of degraded performance for other pathogens.

4.8.2 Organism-Specific Models Exhibit Better Performance than Existing Generalist Models

Comparing the performance of the *RF-OrgSpec* models with the selected predictors (LBtope, iBCE-EL, Bepipred2 & ABCpred) in Figure 8, it is evident that the standard Random Forest model used in this study (even without hyper-parameter tuning or threshold adjustment) consistently outperforms all baseline models from the literature across most performance measures. For all studied organisms, (EBV, HepC & *O.volvulus*) the *RF-OrgSpec* models consistently achieved the highest performance among all predictors, across all performance measures, with only two exceptions: positive predictive value (PPV) for EBV with LBtope and sensitivity (SENS) for *O.volvulus* with Bepipred2. The only predictor that consistently presents performance values comparable to *RF-OrgSpec* is LBtope in the case of the Hepatitis C virus. However, this can be partly attributed to the fact that a portion of the hold-out examples used for evaluating the models are also present in the training data of LBtope (approximately 9.59% of the Hep C hold-out sequences are present in the LBtope training dataset). Similarly, a significant proportion of hold-out Hep C examples are found in the training data of Bepipred2 (16.3%) and iBCE-EL (8.6%). Other predictors are not substantially affected by this data leakage, and it is not observed in the case of the other tested pathogens.

Table 12 presents the performance values and standard errors of all predictors on all test organisms, and shows performance values calculated using only the unseen peptides (not part of the training set) for the case of the Hepatitis C Virus. Out of all organisms studied, only the Hepatitis C Virus had a significant presence of hold-out observations as part of the training sets of some benchmark predictors, with 16.3% of the hold-out sequences present in the training data of Bepipred2, 8.6% in iBCE-EL, and 9.59% in LBtope. For the other organisms, this type of leakage was minimal, and their performance estimates were not affected, allowing the use of the full hold-out set for assessment and presenting a single estimate for their performance. Data leakage appeared to impact the performance of LBtope when evaluated on the hold-out set, as evidenced by a substantial drop in performance observed when using only sequences not present in the training set for validation (Table 12). The values in parentheses in Table 12 indicate the performance variation (up or down) observed when only sequences not present in the training set were used for validation, highlighting the importance of proper data separation to obtain accurate performance estimates.

	Epstein-Barr Virus					
	ACC	AUC	MCC	PPV	NPV	SENS
RF-OrgSpec	0.72 ± 0.01	0.74 ± 0.02	0.32 ± 0.03	0.73 ± 0.02	0.67 ± 0.04	0.92 ± 0.01
RF-Hybrid	0.45 ± 0.02	0.5 ± 0.02	-0.05 ± 0.03	0.64 ± 0.02	0.32 ± 0.02	0.39 ± 0.02
RF-Heter	0.35 ± 0.02	0.36 ± 0.02	-0.25 ± 0.03	0.51 ± 0.03	0.24 ± 0.02	0.3 ± 0.02
ABCpred	0.53 ± 0.02	0.49 ± 0.02	-0.05 ± 0.03	0.65 ± 0.02	0.3 ± 0.03	0.63 ± 0.02
Bepipred2	0.53 ± 0.02	0.42 ± 0.02	-0.11 ± 0.03	0.63 ± 0.02	0.25 ± 0.03	0.7 ± 0.02
iBCE-EL	0.4 ± 0.02	0.46 ± 0.02	0.01 ± 0.03	0.68 ± 0.04	0.34 ± 0.02	0.18 ± 0.02
LBtope	0.55 ± 0.02	0.7 ± 0.02	0.19 ± 0.03	0.78 ± 0.02	0.41 ± 0.02	0.45 ± 0.02
SVMtrip	0.38 ± 0.02	0.43 ± 0.02	-0.09 ± 0.03	0.59 ± 0.03	0.31 ± 0.02	0.21 ± 0.02

	O. volvulus					
	ACC	AUC	MCC	PPV	NPV	SENS
RF-OrgSpec	0.75 ± 0.01	0.83 ± 0.01	0.51 ± 0.02	0.78 ± 0.01	0.73 ± 0.02	0.73 ± 0.02
RF-Hybrid	0.67 ± 0.01	0.75 ± 0.01	0.34 ± 0.02	0.69 ± 0.02	0.66 ± 0.02	0.67 ± 0.02
RF-Heter	0.54 ± 0.01	0.56 ± 0.01	0.06 ± 0.03	0.54 ± 0.02	0.52 ± 0.02	0.67 ± 0.02
ABCpred	0.51 ± 0.01	0.52 ± 0.01	0.02 ± 0.02	0.53 ± 0.02	0.49 ± 0.02	0.58 ± 0.02
Bepipred2	0.63 ± 0.01	0.65 ± 0.01	0.26 ± 0.02	0.61 ± 0.01	0.66 ± 0.02	0.77 ± 0.01
iBCE-EL	0.49 ± 0.01	0.57 ± 0.01	0.02 ± 0.03	0.55 ± 0.04	0.49 ± 0.01	0.1 ± 0.01
LBtope	0.58 ± 0.01	0.59 ± 0.01	0.16 ± 0.02	0.61 ± 0.02	0.55 ± 0.02	0.51 ± 0.02
SVMtrip	0.49 ± 0.01	0.49 ± 0.01	-0.01 ± 0.02	0.51 ± 0.02	0.48 ± 0.01	0.25 ± 0.01

	Hepatitis C Virus					
	ACC	AUC	MCC	PPV	NPV	SENS
RF-OrgSpec	0.75 ± 0.02	0.8 ± 0.02	0.47 ± 0.04	0.67 ± 0.03	0.8 ± 0.02	0.66 ± 0.03 (0.01 ↑)
RF-Hybrid	0.71 ± 0.02	0.75 ± 0.02	0.36 ± 0.04	0.63 ± 0.04	0.74 ± 0.02 (0.01 ↑)	0.53 ± 0.03 (0.01 ↑)
RF-Heter	0.57 ± 0.02	0.55 ± 0.02	0.04 ± 0.04	0.41 ± 0.04	0.63 ± 0.02	0.28 ± 0.03
ABCpred	0.49 ± 0.02	0.58 ± 0.02	0.03 ± 0.04	0.39 ± 0.03	0.64 ± 0.03	0.64 ± 0.03
Bepipred2	0.62 ± 0.02 (0.02 ↓)	0.64 ± 0.02 (0.04 ↓)	0.25 ± 0.04 (0.03 ↓)	0.5 ± 0.03 (0.01 ↑)	0.74 ± 0.03 (0.03 ↓)	0.66 ± 0.03 (0.03 ↓)
iBCE-EL	0.64 ± 0.02 (0.02 ↓)	0.64 ± 0.02	0.14 ± 0.04 (0.02 ↑)	0.62 ± 0.08 (0.10 ↑)	0.64 ± 0.02 (0.02 ↓)	0.12 ± 0.02 (0.02 ↓)
LBtope	0.76 ± 0.02	0.82 ± 0.02 (0.11 ↓)	0.49 ± 0.04 (0.27 ↓)	0.69 ± 0.03 (0.38 ↓)	0.8 ± 0.02 (0.08 ↑)	0.67 ± 0.03 (0.25 ↓)
SVMtrip	0.61 ± 0.02	0.54 ± 0.02 (0.01 ↓)	0.06 ± 0.04 (0.02 ↓)	0.45 ± 0.06 (0.03 ↓)	0.63 ± 0.02	0.14 ± 0.02 (0.01 ↓)

Table 12. Estimates and standard errors of performance for all models tested on the validation sets. Values in parentheses indicate the performance variation (up or down) observed when only sequences not present in the training set were used for validation. Only the Hepatitis C Virus had a significant presence of hold-out observations as part of the training sets of some of the benchmark predictors (16.3% for Bepipred2, 8.6% for iBCE-EL and 9.59% for LBtope). For the other organisms this type of *leakage* was minimal and did not affect the estimates, so the full hold-out set was used for assessment and a single estimate is presented.

The outcomes of this study highlight the significant benefits of adopting organism-specific training for the development of linear B-cell epitope prediction models. Not only did all *RF-OrgSpec* models consistently outperform the *RF-Hybrid* and *RF-Heterogeneous* models but the majority of them also exhibited superior performance compared to the chosen benchmark predictors from the literature, across all assessed performance measures for each respective target pathogen. This validation further supports the notion that leveraging organism-specific data enables the construction of predictive models that capture pathogen-specific patterns, resulting in enhanced predictive capabilities. This work not only expands the understanding of organism-specific modeling but also offers a promising avenue for improving epitope prediction across a range of pathogens.

4.8.3 Illustrative Example: *Onchocerca volvulus* Results

This section offers a deeper exploration of the predictions obtained for the *O. volvulus* dataset to provide a comprehensive illustration of the organism-specific modelling results. Figure 10 displays the receiver operating characteristic (ROC) curves obtained for all predictors on the hold-out datasets for each target pathogen. The right panel of Figure 10 shows the ROC curves for all predictors on the *O. volvulus* hold-out dataset. This graph clearly demonstrates significant performance gains with the organism-specific model. The *RF-OrgSpec* model exhibited excellent robustness to different threshold values (AUC = 0.83). Additionally, the *RF-Hybrid*, which also incorporated organism-specific data in its training, showed reasonably good performance (AUC = 0.75). These findings further support the efficacy of organism-specific training in improving predictive capabilities for epitope prediction.

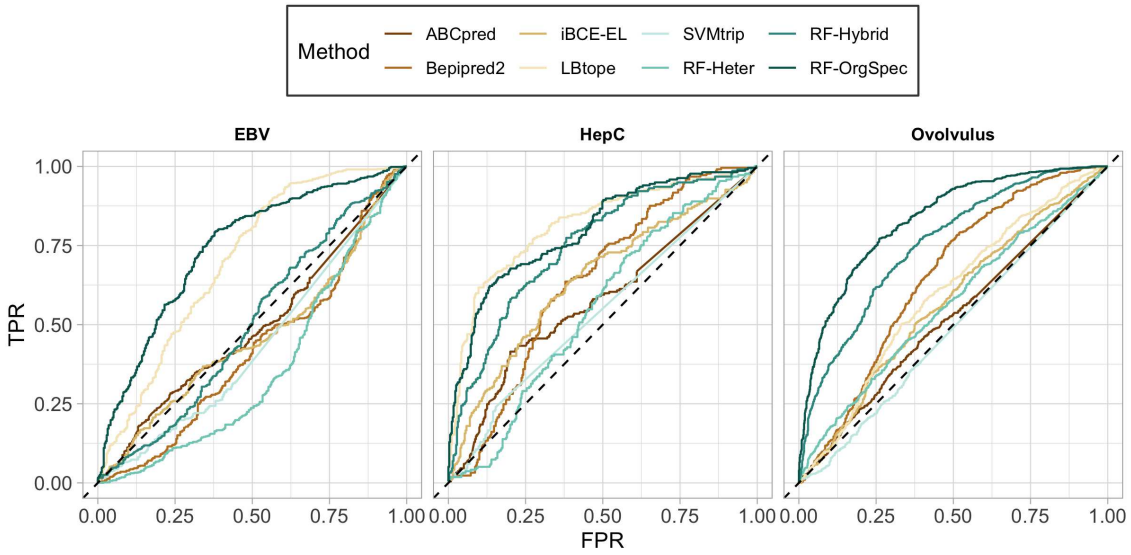


Figure 10. ROC curves for all predictors tested on the hold out sets for the Epstein-Barr Virus (left), Hepatitis C Virus (centre) and *Onchocerca volvulus* (right).

Figures 11 and 12 illustrate the regions predicted as epitopes by the organism-specific models for the 22 hold-out proteins of the *O. volvulus* hold-out dataset, using the default threshold value of 0.5. These results showcase not only the excellent agreement of the *RF-OrgSpec* predictions with the known epitope labels on the hold-out proteins, but also highlight a number of newly identified potential epitopes that may exist in those proteins. The peptides output by the *O. volvulus* model with an average probability of over 0.75 are listed in Table 13. These results provide valuable insights into the effectiveness and potential of the organism-specific approach for epitope prediction.

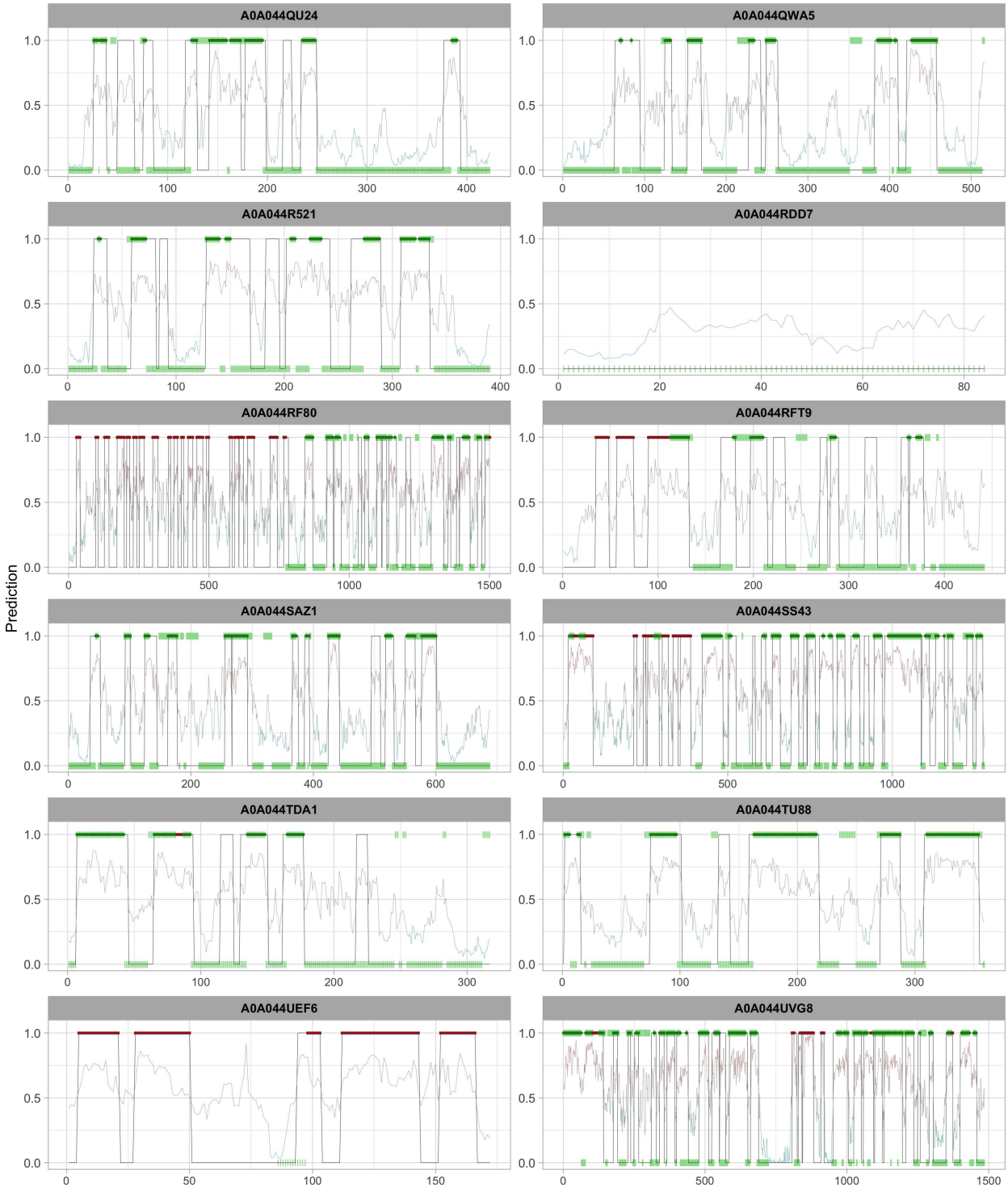


Figure 11. *RF-OrgSpec* predictions for the proteins on the *O. volvulus* hold-out set (Part 1). The narrow line shows the predicted probability returned by *RF-OrgSpec*, which was thresholded to yield positive/negative predictions. Light-green highlights indicate regions that were labelled as positive/negative in the IEDB data. Narrow dark green indicates **true positive** predictions, and red indicates new candidate targets, that is, regions without known labels that were predicted as positive by the *RF-OrgSpec* model.

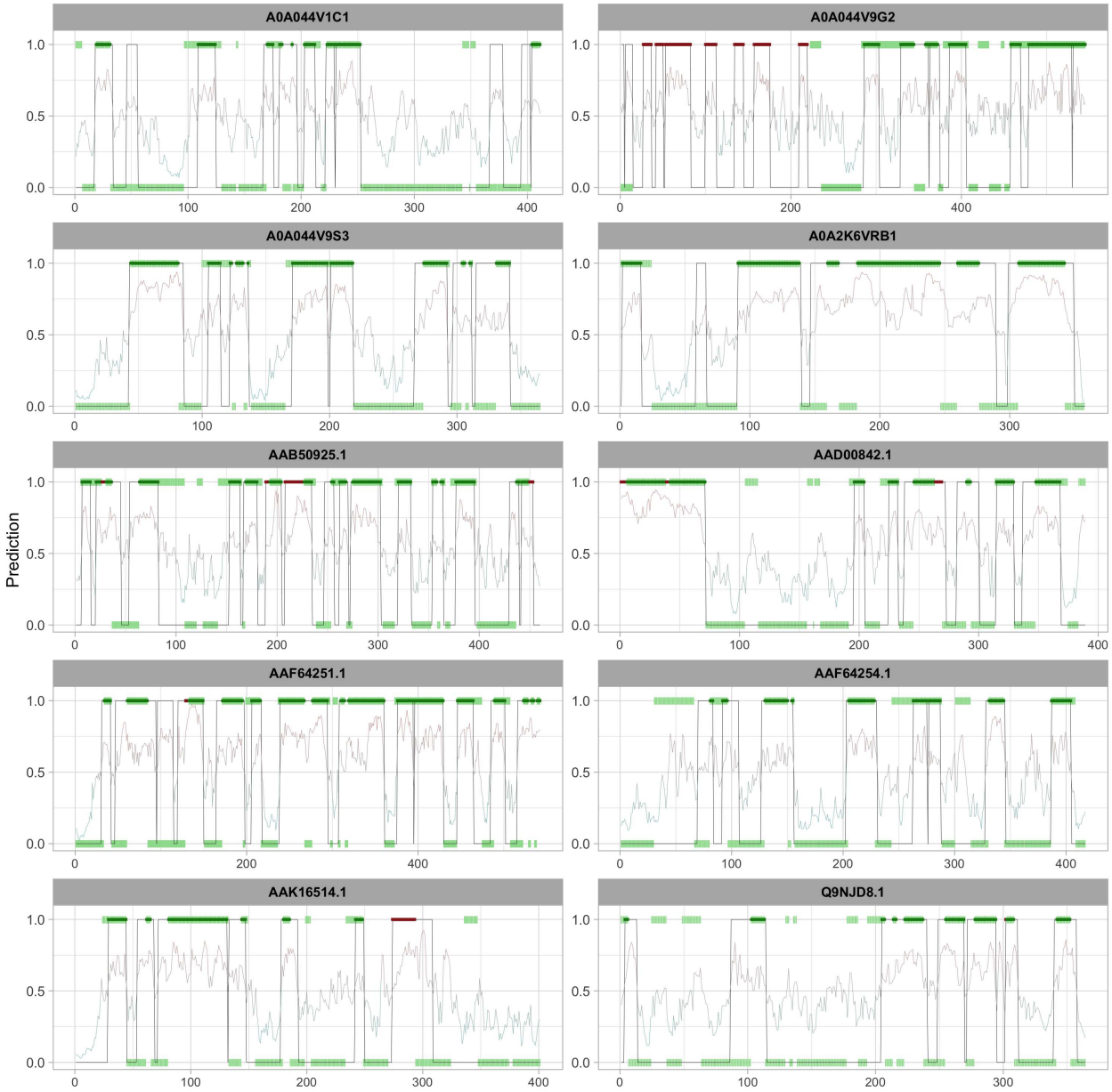


Figure 12. *RF-OrgSpec* predictions for the proteins on the *O. volvulus* hold-out set (Part 2). The narrow line shows the predicted probability returned by *RF-OrgSpec*, which was thresholded to yield positive/negative predictions. Light-green highlights indicate regions that were labelled as positive/negative in the IEDB data. Narrow dark green indicates **true positive** predictions, and red indicates new candidate targets, that is, regions without known labels that were predicted as positive by the *RF-OrgSpec* model.

Table 13. Predicted target regions in proteins of the *O. volvulus* hold-out set with average predicted probabilities above 0.75. Column *'Matches'* indicates IEDB epitope IDs for which the sequence identity returned a match of over 80%. Two candidate epitopes have no match on IEDB (Matches = **N/A**), and may represent new targets for experimental investigation.

Protein	**Start**	**End**	**Length**	**Prob.**	**Matches**
A0A044UVG8	845	882	38	0.86	852235 (100%); 855900 (100%)
AAD00842.1	1	71	71	0.83	851260 (100%); 856295 (100%); 853774 (100%); 854736 (100%); 854918 (93.333%)
A0A044V9S3	43	85	43	0.81	854970 (100%); 851810 (100%); 852047 (100%); 857301 (100%); 857276 (100%); 851449 (87.5%); 854446 (85.714%)
A0A044SS43	15	90	76	0.80	854918 (100%); 851260 (100%); 856295 (100%); 853774 (100%); 854736 (100%); 856311 (85.714%)
AAF64251.1	120	149	30	0.80	854055 (100%); 853779 (100%); 852831 (93.333%)
A0A2K6VRB1	299	349	51	0.80	855140 (100%); 855121 (100%); 855119 (100%); 851603 (100%); 855456 (100%); 853564 (90%)
AAF64251.1	376	394	19	0.80	854291 (100%); 853775 (100%); 851576 (100%)
A0A044SAZ1	425	442	18	0.80	855469 (100%); 855469 (100%); 854673 (100%); 855208 (100%); 945036 (83.333%); 945545 (83.333%)
A0A044SS43	875	901	27	0.80	856192 (100%)
A0A044SAZ1	577	601	25	0.79	851740 (100%)
A0A044UVG8	829	841	13	0.79	**N/A**
A0A044SS43	945	970	26	0.79	853848 (100%); 856088 (83.333%)
A0A044SS43	743	766	24	0.78	853390 (100%); 854905 (100%); 854712 (100%)
A0A044SS43	977	1087	111	0.78	856569 (100%); 855321 (100%); 852138 (100%); 856604 (100%); 851924 (100%)
AAF64251.1	516	542	27	0.78	856224 (100%)
A0A044SS43	779	819	41	0.77	856056 (100%)
A0A044RF80	1485	1501	17	0.77	854139 (100%); 851875 (100%)

Table 13 – continued from previous page

Protein	Start	End	Length	Prob.	Matches
A0A044SS43	636	659	24	0.77	851949 (100%); 854897 (100%); 852938 (83.333%)
A0A044SS43	423	485	63	0.77	853828 (100%); 852292 (100%); 853331 (100%); 854720 (100%); 853875 (100%); 851706 (100%)
A0A044UVG8	1	141	141	0.77	853160 (100%); 851567 (100%); 854638 (100%); 852144 (100%); 853634 (100%); 854011 (100%); 854765 (100%); 853789 (100%); 855619 (100%)
A0A044UVG8	1400	1458	59	0.77	854721 (100%); 855565 (100%)
A0A044UVG8	1059	1094	36	0.77	855021 (100%); 856367 (100%)
AAF64254.1	328	345	18	0.77	854085 (100%)
A0A044UVG8	661	688	28	0.77	854757 (100%); 856403 (100%)
A0A044TU88	308	354	47	0.76	852155 (100%); 851711 (100%); 853615 (100%)
A0A044SS43	915	929	15	0.76	**N/A**
A0A044QWA5	421	457	37	0.75	854962 (100%); 851542 (100%); 851715 (100%); 855420 (100%)
AAF64251.1	239	294	56	0.75	856870 (100%); 853605 (100%); 852659 (100%); 851234 (100%)

This series of results exemplifies how the heightened overall performance of organism-specific models, in contrast to state-of-the-art predictors, can be immensely valuable in advancing the identification and selection of epitopes for diagnostic targets and vaccine candidates for infectious diseases. Notably, the enhanced positive predictive values (PPV), as indicated in Figure 8, indicate that the predicted targets hold a high likelihood of being antigenic, thereby enhancing the efficiency of epitope discovery processes facilitated by the proposed organism-specific models. This could be a consequence of unique, idiosyncratic epitope patterns specific to different species, that would be overlooked by generalist predictors. For this reason, organism-specific models hold exceptional relevance for pathogen types typically underrepresented in broad, generic epitope training databases.

4.9 Insights from Organism-Specific Epitope Prediction Results

The results described in this section indicate a clear improvement in performance resulting from the use of organism-specific models, when compared to generalist predictors trained on heterogeneous, or even hybrid, data. While a comprehensive analysis of the underlying factors contributing to these performance disparities is beyond the purview of this study, several potential non-mutually exclusive hypotheses can be suggested for further exploration.

4.9.1 Exploring Feature Relevance and Implications in the Context of Organism-Specific Modelling

In an attempt to unravel the possible factors contributing to the superior performance exhibited by organism-specific data-trained models, a visual exploration of feature relevance was conducted. This section of the study delves into the details of this investigation. Throughout these figures, the various families of features are denoted using the following nomenclature:

- **AAdescr.**: Amino Acid Descriptors (66 features), 66 physiochemical descriptors (Cruciani properties [115], Kidera factors [116], Z-scales [117], FASGAI indices [118], T-scales [119], VHSE-scales [120], ProtFP descriptors [121], ST-scales [122], BLOSUM indices [123] & MS-WHIM scores [124]).
- **Atoms**: Counts of Carbon, Hydrogen, Nitrogen, Oxygen and Sulphur atoms (5 features).
- **CT**: 343 Conjoint Triad frequencies [85] (343 features).
- **Entropy**: Sequence entropy/information entropy [113] (1 feature).
- **Freq-1AA**: Frequencies of occurrence of each amino-acid (20 features).
- **Freq-2AA**: Frequencies of occurrence of each dipeptide (400 features).
- **Freq-Types**: Frequencies of occurrence of each of the 9 amino-acid types (aliphatic, acidic, polar, etc.) (9 features).
- **Mol. weight**: Molecular weight (1 feature).

All features were derived from the local 15 amino acid-wide neighbourhood surrounding each position within a protein sequence (as described in Section 2.6.1). Figures 13, 14 & 15 offer an insightful analysis into the significance of distinct features and feature groups, providing valuable insights into their contributions to the predictive capabilities of both *OrgSpec* and *Heterogeneous* models.

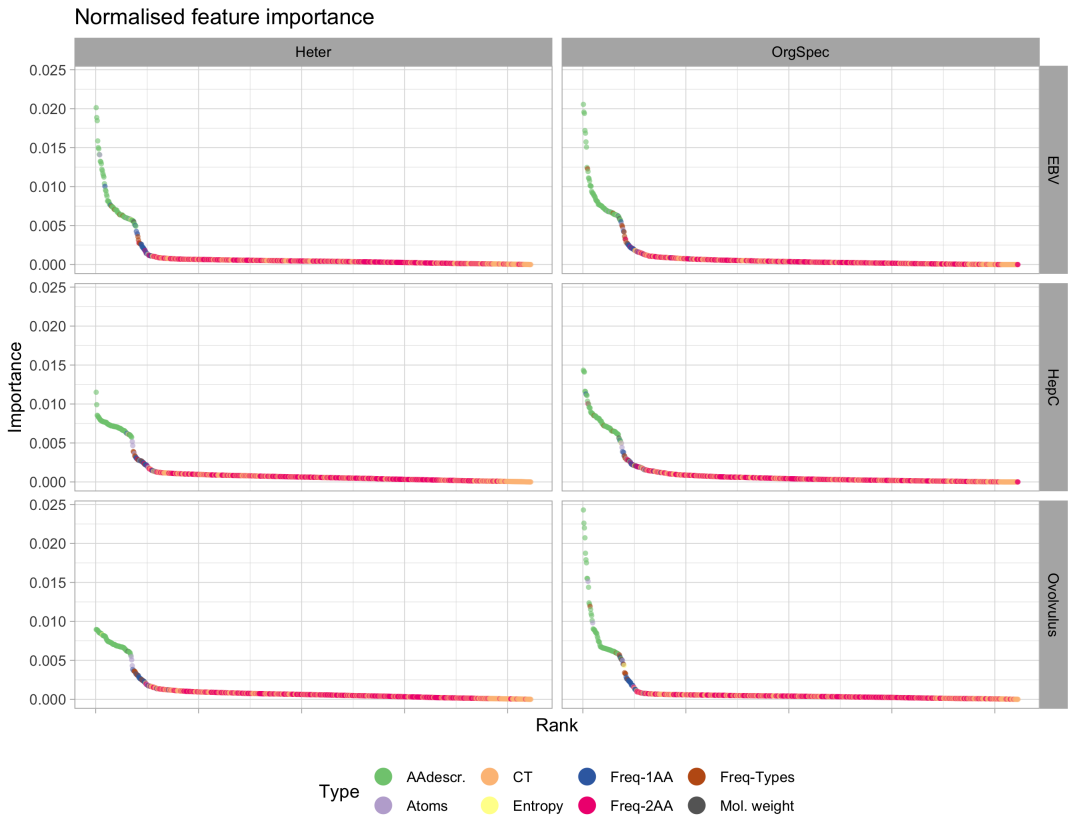


Figure 13. Feature importance of *organism-specific* (OrgSpec) and *heterogeneous* (Heter) RF models coloured by feature family. In all cases, the *Freq-2AA* and *CT* features contribute very little to the predictive ability of the RF models, and the *AAdescr*-type features are consistently selected as the most relevant by Random Forest.

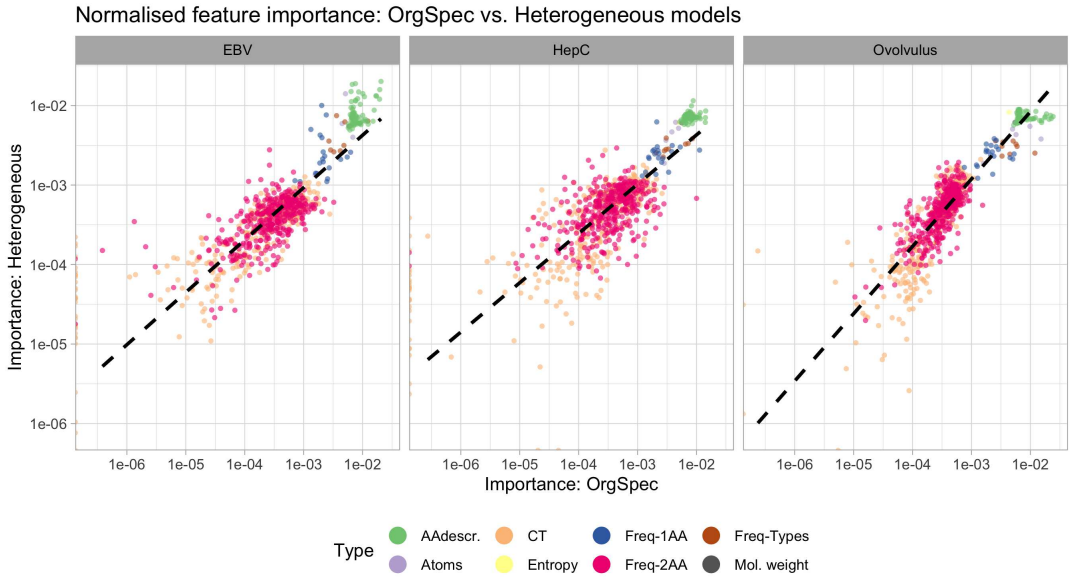


Figure 14. Comparison of feature importance between *organism-specific* and *heterogeneous* data-trained models stratified by pathogen. The dashed regression line serves as a reference to qualitatively assess the strong correlation between the two sets. Note: both axes are log-scaled.

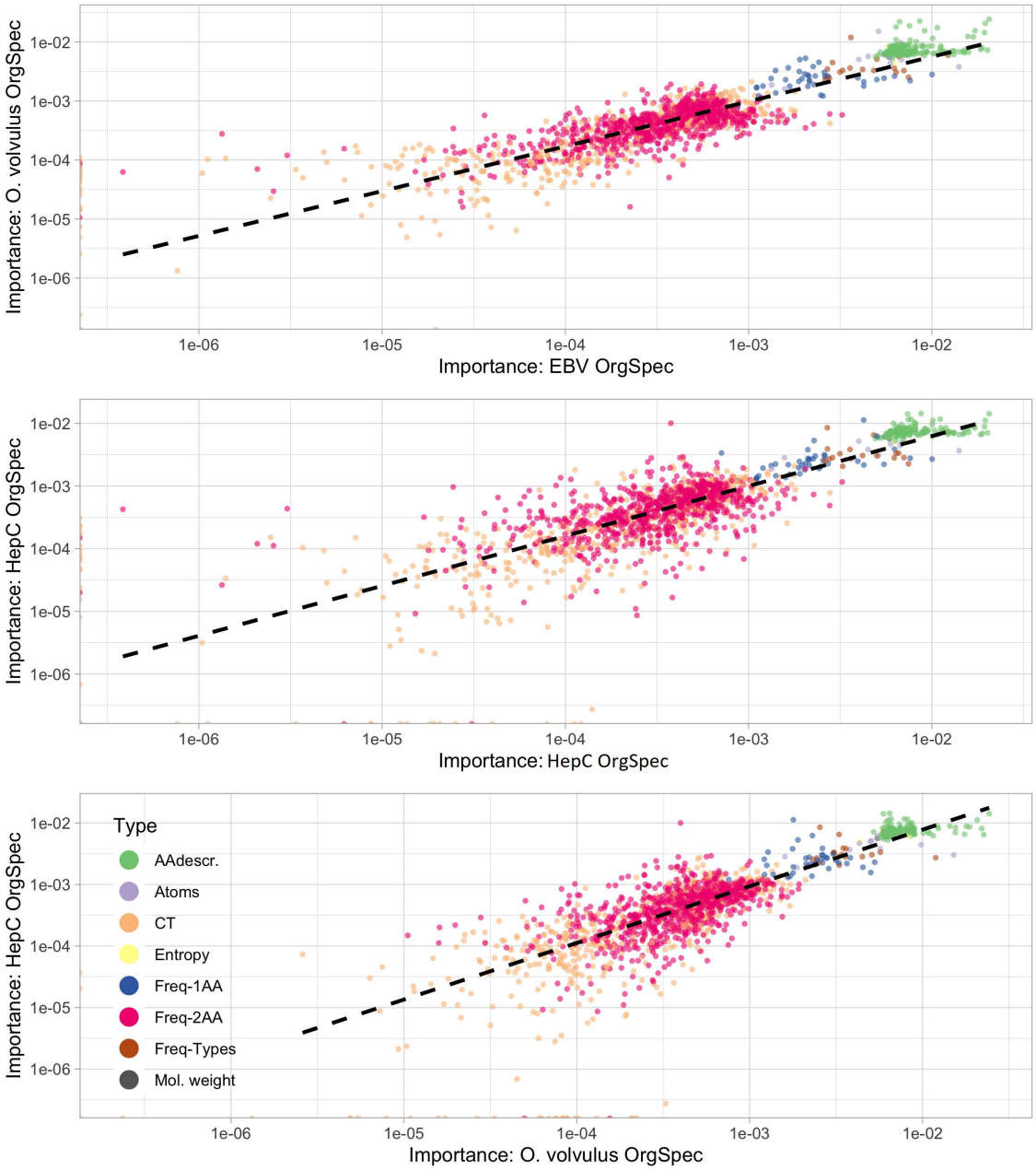


Figure 15. Feature importance for different *organism-specific* models. The dashed regression line provides a reference for qualitatively assessing the high correlation. Note: both axes are log-scaled. This again highlights the high relevance of the physiochemical descriptors *AAdescr*, followed by the *atom* counts and single-amino acid frequencies (*Freq-1AA*). Dipeptide (*Freq-2AA*) and *CT* frequencies do not contribute significantly to any of the models.

From this analysis, there are several interesting general insights that can be derived regarding which feature groups contribute the most to the predictive ability of both *organism-specific* (OrgSpec) and *heterogeneous* (Heter) models (Figures 13, 14 & 15). One notable observation is the disproportionately large prevalence of *AA descriptors*-type features among the most influential for predictive accuracy. Another is the apparent irrelevance of di-peptide frequencies or Conjoint Triads in the context of the linear B-cell epitope prediction problem modelled here.

All three figures (13, 14 & 15) display the same patterns of relative importance of feature families: highlighting the high relevance of the physiochemical descriptors (*AAdescr*), followed by the atom counts (*Atoms*) and single amino-acid frequencies (*Freq-1AA*). Dipeptide (*Freq-2AA*) and *CT* frequencies exhibit minimum contribution to the predictive capabilities of any of the models. However, for the context of this study, it's especially valuable to focus on features that consistently exhibit higher relevance for organism-specific (OrgSpec) models compared to heterogeneous (Heterogeneous) ones.

As indicated in Figures 16 & 17 the BLOSUM1 feature [162] stands out prominently as highly relevant overall, with particular significance for the organism-specific models. The BLOSUM1 feature is strongly correlated with hydrophobicity, boasting an r^2 value of 0.94 according to [162]. In the windowed data representation used in this study, this feature quantifies the average hydrophobicity of the 15-amino-acid neighborhood surrounding a specific position on the protein. Hydrophobicity and hydrophilicity are closely linked to epitope accessibility within the protein structure. Hydrophilic polar regions are typically found on the protein's surface, continually exposed to antibodies, whereas hydrophobic regions often engage in interactions within the protein's core or with other cellular components, rendering them less accessible to the serological immune response [35]. Additionally, there are other features that consistently emerge as highly relevant, although not as prominently as BLOSUM1. These include features like ProtFP1, Z1, VHSE8, and F5 . These features are composite scales derived from algebraic transformations of underlying physicochemical properties, lacking the same direct interpretability as BLOSUM1, which hinders the formulation of biochemical hypotheses based on them. While intriguing, the detailed investigation of underlying mechanisms potentially represented by these particular features is beyond the scope of this work.

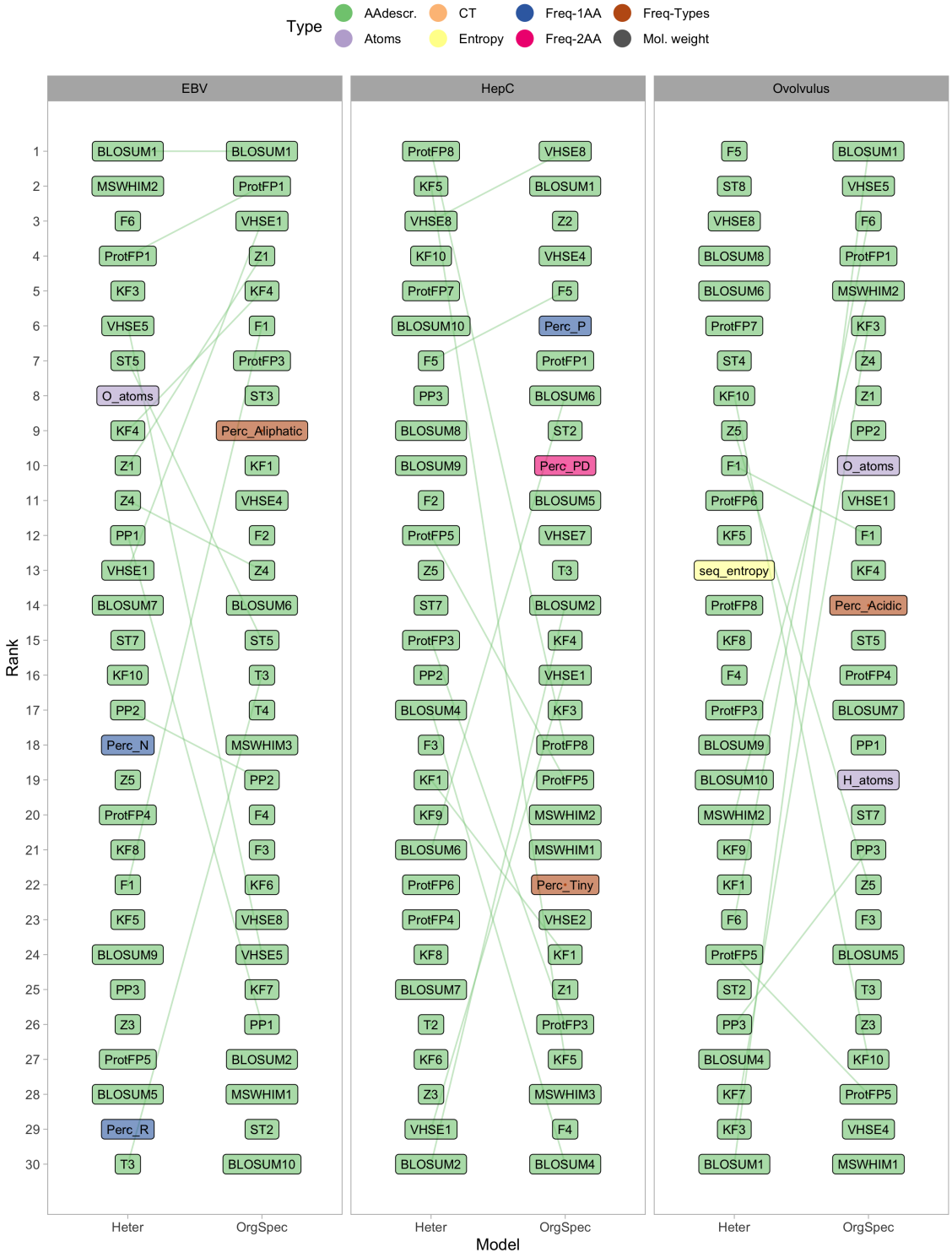


Figure 16. Thirty most highly relevant features according to each Random Forest model trained for each pathogen (organism-specific and heterogeneous). The high prevalence of *AAdescr* features is clear, although they make up only a fraction of the total feature space (66/845 features $\approxeq$ 7.8%).

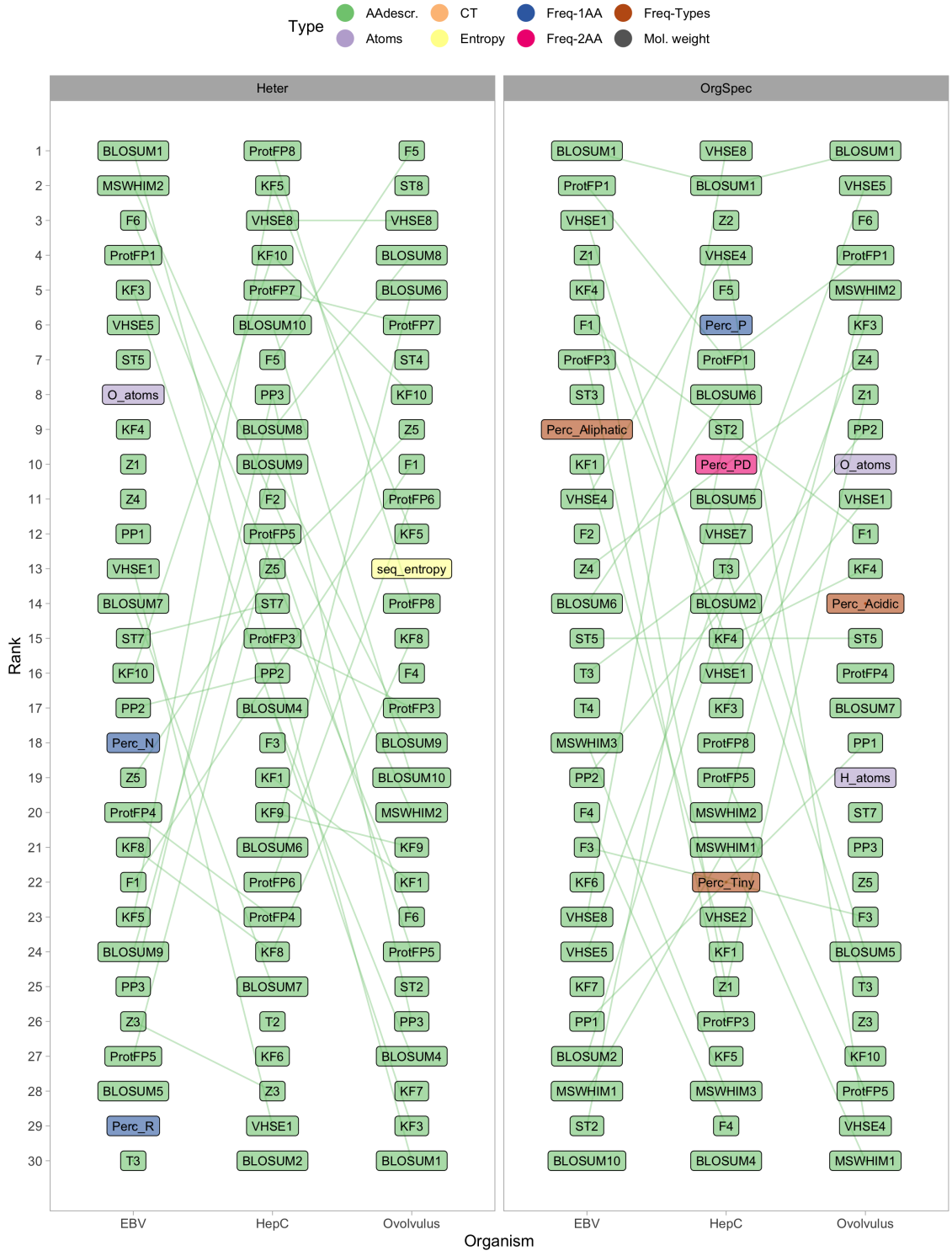


Figure 17. An alternate representation of the top thirty most relevant features selected by the Random Forest models. The BLOSUM1 feature appears as particularly relevant for all organism-specific models.

4.9.2 Spatial Distribution of Observations

Another factor that could offer insight into the improved performance of organism-specific models is the potential variance in the spatial distribution of epitopes within the feature space, depending on the pathogen. To explore the local data structure, t-distributed stochastic neighbor embedding (t-SNE) projections [163] were utilised to investigate whether data originating from different pathogens exhibit distinct clustering or neighborhood patterns concerning positive and negative observations. Figure 18 presents these t-SNE projections.

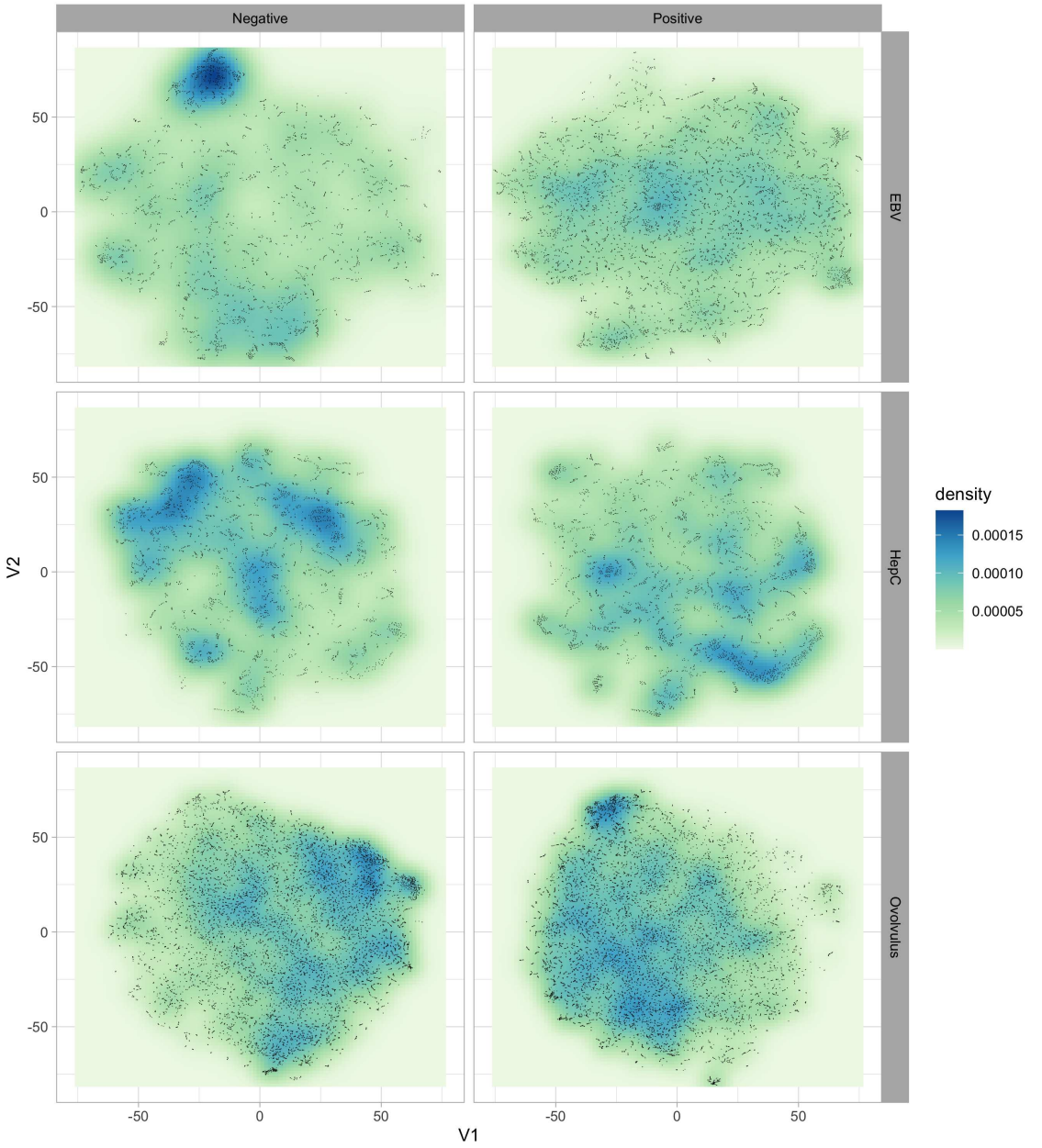


Figure 18. Estimated probability density of epitope and non-epitope observations in the t-SNE projection. Clear distinct regions of high density of positive/negative observations can be seen which occupy different portions of the feature space. Epitopes (*positive*) of different pathogens tend to occur in distinct regions of the feature space. Models trained on combined (heterogeneous) data would not be able to explore these patterns.

Figure 18 illustrates the estimated density of observations on the 2D t-SNE projection of the data, categorised by pathogen and class. The V1-V2 coordinates remain consistent across all panels and clear distinct regions with high densities of both positive and negative observations can be seen, which occupy different portions of the feature space. The figure clearly demonstrates how the density of positive and negative examples not only varies depending on the pathogen but also reveals that regions with a high density of positive examples for one organism may simultaneously contain high densities of negative examples for others. For instance, the high-density top-left portions of the negative examples for both EBV and HepC coincide with a corresponding high-density region of positive examples for *O. volvulus*.

This type of pattern can help explain the effectiveness of organism-specific training from a data mining perspective, although not necessarily a biological one. Generalist models trained on heterogeneous data might struggle to discern these organism-specific patterns since they would appear as a more mixed combination of positive and negative examples if data from multiple pathogens were merged into a single training set. This could in effect prevent those models from identifying promising regions of the feature space that were potentially rich in epitopes of a specific pathogen, leading to reduced predictive performance.

4.10 Organism-Specific Modelling for Epitope Prediction Conclusions

This study investigated the use of organism-specific data for improving the performance of linear B-cell epitope prediction. Organism-specific Random Forest models were developed for three distinct pathogens, namely the Epstein-Barr Virus, Hepatitis C Virus, and the roundworm *Onchocerca volvulus*. Our findings revealed that these models presented substantial performance improvements in comparison to similar models trained on heterogeneous and hybrid datasets, across multiple key performance indicators. These results suggest that carefully pre-selecting the most relevant data and training bespoke models for specific pathogens is preferable to the common strategy of increasing and diversifying the training dataset.

Furthermore, performance comparisons also highlight that this organism-specific modeling approach can yield results that are not only on par with but, in several instances, surpass those of common predictors from the literature; despite the fact that the predictors developed in this study were relatively straightforward proof-of-concept models without extensive fine-tuning. Additionally, only basic features derived from the amino acid sequence itself were utilised, with no intricate feature engineering applied. Further enhancements to organism-specific predictors, including model refinement and the incorporation of more

informative features, have the potential to elevate predictive performance even further. While these results do not diminish the relevance of generalist predictors, which remain invaluable for the investigation of pathogens for which little or no specific data is available - they certainly propose a potent and readily applicable new approach for researchers dealing with relatively data-abundant organisms.

5. Exploring the Limits of Organism-Specific Training for Linear B-Cell Epitope Prediction

The previous chapter (Chapter 4, Organism-Specific Modelling for Epitope Prediction) showcased the effectiveness of organism-specific training in improving linear B-cell epitope prediction performance for organisms with abundant epitope data resources (data-rich organisms). This chapter investigates the limits of organism-specific training for LBC epitope prediction, by systematically quantifying the effect of the amount of training data on the performance of the models developed. The findings from this investigation reveal that even models trained on modest-sized organism-specific datasets can outperform comparable models trained on larger, heterogeneous and mixed datasets. Furthermore, these models exhibit superior performance compared to widely-used well-established predictors from the scientific literature (which are trained on heterogeneous data). These results indicate potential for a much broader application of pathogen-specific models for the prediction of linear B-cell epitopes, which may facilitate the study of data-poor organisms such as emerging or neglected pathogens.

5.1 The Scarcity of Epitope Prediction Data

As mentioned previously, the current prevailing approach for epitope prediction involves training predictive models on extensive heterogeneous datasets encompassing observations from various organisms such as, prokaryotes, viruses, fungi, protozoan, humans and other eukaryotes (Chapter 4, Section 4.1, Table 5). However, this work (Chapter 5) has shown that training models with smaller, organism-specific datasets can improve epitope prediction performance, for data-rich organisms. The previous chapter (4) investigated organism-specific training for three distinct organisms: the nematode *Onchocerca volvulus*, Epstein-Barr Virus, and Hepatitis C Virus. The selection of these organisms was driven by the availability of an ample volume of validated epitope observations, both positive and negative, within the Immune Epitope Database (IEDB) [28]. Unfortunately, for most organisms, substantial amounts of validated epitope data are not available; several factors may be contributing to this scarcity: the organism might be associated with an emerging or neglected disease that has not garnered substantial research focus. The organism might possess only a limited number of epitopes. In cases where the organism's impact on human health is negligible, research efforts such as, immunological investigations and epitope data collection might be scant. Moreover, the experimental validation of epitopes demands significant resources and time, further compounding the challenge.

5.2 Outlining the Limits of Organism-Specific Training Investigation

This chapter seeks to explore the viability of organism-specific training for organisms with relatively limited available epitope data. It aims to probe the limits of organism-specific training, addressing two key questions:

i How does the quantity of available organism-specific training peptides impact prediction performance?
ii What is the minimal organism-specific data required to achieve superior model performance compared to models trained on extensive, diverse datasets?

To address these inquiries, we assess and compare predictive performance of models trained on reduced training sets, mixed data, and large, heterogeneous datasets. Additionally, we contrast these outcomes with those of four generalist predictors from the scientific literature: Bepipred2.0 [63], LBtope [57], iBCE-EL [39], and ABCpred [14], across diverse performance indicators.

5.3 The Limits of Organism-Specific Training Investigation Methods

5.3.1 Dataset Generation

For this investigation, the same datasets from Chapter 4 were used. Data from three distinct pathogens, namely *Onchocerca volvulus* (taxonomy ID: 6282), Epstein-Barr Virus (taxonomy ID: 10376), and Hepatitis C Virus (taxonomy ID: 11102), were utilised, the generation of these datasets was detailed in Section 4.4.2. Segmentation occurred at the protein level, with entries belonging to the same protein or from proteins displaying over 80% sequence coverage and similarity being grouped together in the same split. This investigation utilised four distinct datasets: the *Hold-out* set (comprising approximately 25% of the organism's epitope data), the *Organism-specific* set (consisting of the remaining 75% of the organism's epitope data), the *Heterogeneous* dataset (excluding any observations from the target pathogen), and the *Hybrid* dataset (combining both organism-specific and heterogeneous data) for each respective target pathogen.

The objectives of this investigation were to explore the influence of the size of organism-specific training datasets on linear B-cell epitope prediction performance, while also aiming to determine approximate lower bounds for the amount of data necessary for effective organism-specific training as a viable alternative to models developed on larger, heterogeneous datasets. To achieve this, several heterogeneous, hybrid and *reduced*

organism-specific training datasets were generated for each target organism; the organism-specific datasets were based on the available organism-specific model development data described above. For each organism and each desired training set size, the full organism-specific model development data was split into smaller non-overlapping *Organism-specific* data sets, each containing data from between 20 and 500 peptides . The same class balance as the full organism-specific dataset was maintained in all reduced subsets.

For each target organism, multiple *reduced* organism-specific datasets were generated. The total number of available organism-specific training peptides (4819 for *O. volvulus*, 1702 for Hep C and 2557 for EBV) determined the number of replicates produced for each *reduced* dataset size. Table 14 presents the number of replicates corresponding to each reduced dataset size (set size = number of organism-specific peptides in the set). The distribution of positive (+) and negative (-) peptides in each organism-specific dataset (training and hold-out sets) for the target pathogens is also shown in Table 14.

	O. volvulus	Hepatitis C Virus	Epstein-Barr Virus
Hold-out peptides	$(832+ / 777-)$	$(218+ / 358-)$	$(625+ / 315-)$
Train/Model dev. peptides	$(2441+ / 2378-)$	$(919+ / 783-)$	$(1746+ / 811-)$
20-peptide sets (N_{20})	237	83	124
40-peptide sets (N_{40})	118	41	62
60-peptide sets (N_{60})	79	27	42
80-peptide sets (N_{80})	59	21	31
100-peptide sets (N_{100})	47	17	25
150-peptide sets (N_{150})	32	11	16
200-peptide sets (N_{200})	24	8	12
250-peptide sets (N_{250})	19	6	10
300-peptide sets (N_{300})	16	5	8
400-peptide sets (N_{400})	12	4	6
500-peptide sets (N_{500})	9	4	5

Table 14. Summary of organism-specific datasets: number of positive (+) / negative (-) peptides in each set, and number of replicates for each reduced dataset size. I.e. for the 20-peptide sets (N_{20}) there are 237 *O. volvulus* datasets (each made up of 20 organism-specific peptides), 83 Hep C datasets and 124 EBV sets.

Table 14 provides an overview of the *reduced* organism-specific datasets generated for each target pathogen. Using these variable-sized organism-specific training datasets, two categories of *hybrid* datasets were also assembled: *Hybrid-A* and *Hybrid-B*.

- *Hybrid-A*, consisting of the organism-specific peptides (at different sizes) plus an equal number of peptides randomly sampled from other pathogens. *Hybrid-A* datasets therefore comprised twice the number of peptides as their corresponding organism-specific counterparts, maintaining a balanced distribution of $50 - 50\%$ between organism-specific and 'other' peptides.

- *Hybrid-B*, comprising the organism-specific peptides plus additional peptides sampled from other pathogens to achieve a fixed dataset size of 1000 training data peptides (e.g., 20 organism-specific + 980 'other' peptides; 40 organism-specific + 960 'other', and so forth). *Hybrid-B* datasets maintained a fixed size (1000 peptides) while varying the balance between data from the target pathogen (organism-specific data) and data from other organisms.

For each target organism, an additional category of datasets were also assembled: the *Heterogeneous* datasets. Heterogeneous data was collated by randomly sampling observations grouped by taxonomy ID from the full IEDB export, excluding any observations associated with the specific target organism. For each organism, 6000 labeled peptides exclusively from non-target organisms were extracted, maintaining a balanced class distribution of 50% epitope and 50% non-epitope peptides, to form a large *heterogeneous dataset*. From this heterogeneous data several smaller fixed-size *heterogeneous* datasets were also extracted by sub-sampling (without replacement) 1000 (non-target organism) peptides from the heterogeneous data, 30 replicates were extracted for each organism.

5.3.2 Experimental Protocol Overview

Figure 19 provides an overview of the experimental protocol for assessing the limits of organism-specific model training for linear B-cell epitope prediction. This figure shows that for each target pathogen:

- The organism-specific (target pathogen data) is first split into training (model development - 75%) and hold-out sets (25%).
- **(A)** For each desired organism-specific dataset size (number of target pathogen peptides), the organism-specific model development data is partitioned into non-overlapping subsets of the desired size, each preserving the original class balance distribution.
- **(B)** Two categories of hybrid datasets are constructed using the reduced organism-specific data replicates: *Hybrid-A* maintains a fixed 50-50 balance between organism-specific and heterogeneous data at all dataset sizes; *Hybrid-B* adds the required number of non-target organism observations to reach a dataset size of 1,000 peptides, resulting in datasets with varying proportions of organism-specific VS 'other' peptides.
- **(C)** Baseline datasets comprising 1000 non-target pathogen peptides are generated through diverse sub-sampling (without replacement) from the heterogeneous data.
- All datasets are then employed to train Random Forest models, which have their performance assessed on organism-specific hold-out data.

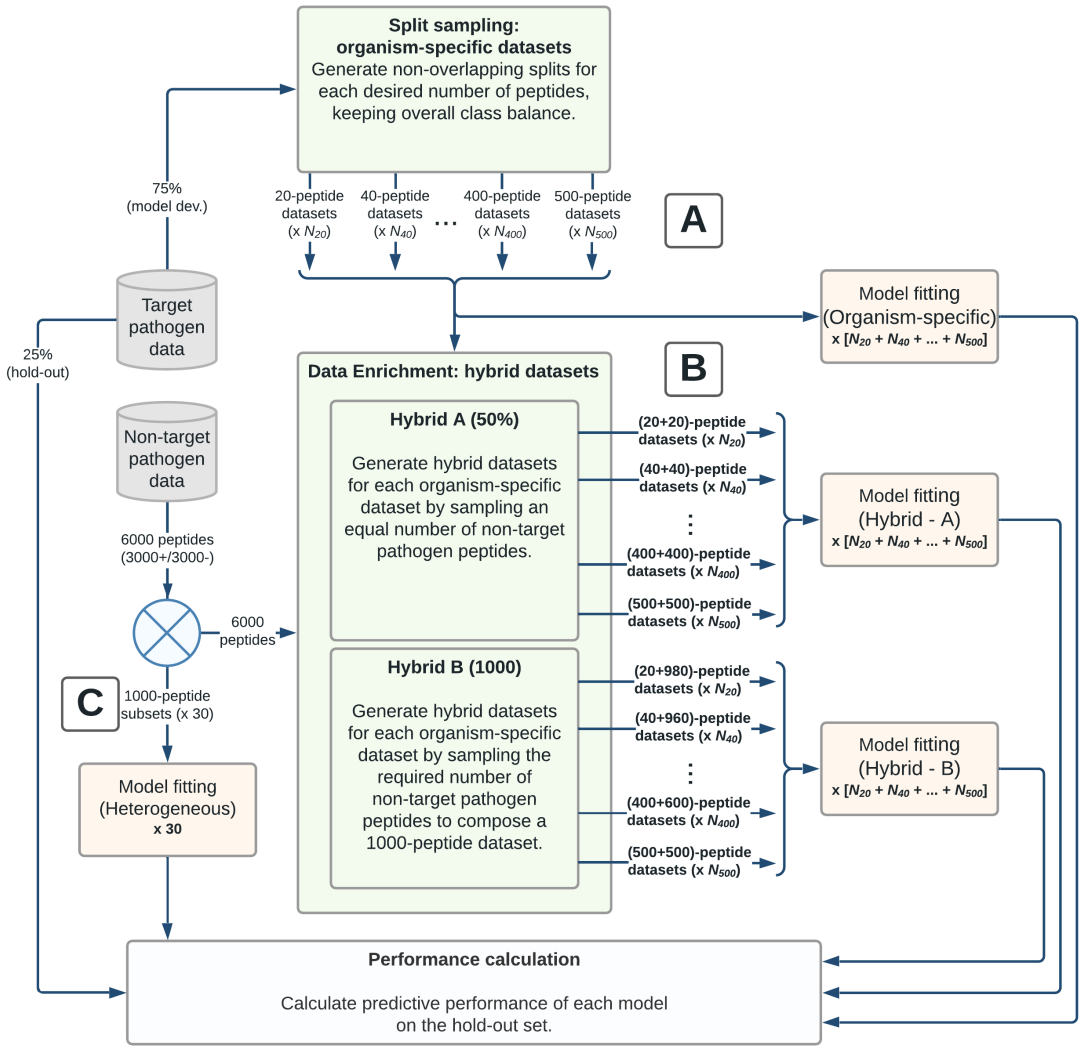


Figure 19. Experimental protocol for testing the limits of organism-specific model training for linear B-cell epitope prediction. **(A)** For each pathogen and each desired data size, the model development data set is split into non-overlapping subsets, each maintaining the original class balance of the data. **(B)** Two sets of hybrid data sets are composed based on the organism-specific reduced-data replicates: *Hybrid-A* maintains a fixed 50-50 balance between organism-specific and heterogeneous data at all data set sizes; *Hybrid-B* adds the required number of non-target organism observations to complete a data set of 1,000 peptides, resulting in sets with variable proportions of organism-specific peptides. **(C)** Baseline data sets composed of 1,000 exclusively non-target pathogen peptides are also generated based on different sub-samplings (without replacement) from the heterogeneous data. All data sets are used to train Random Forest models, which then have their performance assessed on organism-specific hold-out data.

5.4 Modelling and Performance Assessment

Several epitope prediction models were constructed through the training of Random Forest (RF) predictors on the aforementioned training datasets. For each target organism, RF models were trained on:

- *'OS-full'*: The full organism-specific model development training dataset.
- *'Organism-Specific'*: Several reduced organism-specific training datasets, relating to 11 size categories (from 20-peptide sized datasets to 500-peptide sized datasets).
- *'Hybrid-A'*: The same number of organism-specific peptides as other (non-target organism) peptides, for each *organism-specific* size category.
- *'Hybrid-B'*: The *organism-specific* datasets plus a number of other, non-target peptides to reach a dataset size of 1000 peptides.
- *'Heterogeneous 1K'*: 30 lots of 1000 non-target organism peptides. (These datasets are the same size as all *Hybrid-B* datasets and can be thought of as *Hybrid B* datasets with 0 % organism-specific data and 100 % non-target peptide data.)
- *'Heter 6K'*: A large heterogeneous dataset consisting of 6000 non-target pathogen peptides.
- *'OS-full + 6K'*: A large hybrid dataset combining the full organism-specific training dataset (*OS-full*) and the large heterogeneous dataset (*Heter 6K*).

The RF implementation utilised Scikit-learn version 0.24.1 [154] with default hyperparameter values. The selection of Random Forest was based on preliminary experimentation as detailed in Section 4.6.1 fostering direct comparability with prior research outcomes (Chapter 4). The trained models were employed to generate predictions for the organism-specific hold-out datasets. Evaluation of prediction performance encompassed several distinct measures: Balanced Accuracy (BAL.ACC), Matthew's Correlation Coefficient (MCC), Area Under the ROC Curve (AUC), Positive Predictive Value (PPV), Negative Predictive Value (NPV), and Sensitivity (SENS). Given that these performance indicators were computed on the hold-out datasets (which remained unseen by the models throughout development except during testing), it can be assumed that the resulting values are reasonably indicative of the models' generalisation capabilities in epitope prediction for proteins derived from the respective target pathogens.

The mean estimated performance and corresponding standard errors for each quality measure were computed from the replicates across the various pathogen types and dataset sizes. These findings were then juxtaposed against a series of baseline benchmarks including: the observed performance of Bepipred2.0 [63], LBtope [57], iBCE-EL [39] and ABCpred [14] on the hold-out set of each pathogen; and the results from the *OS-full*, *Heter 6K* and *OS-full + 6K* models on the hold-out datasets.

5.5 Results

Figures 20 and 21 present the mean performance results of each set of models on the respective pathogen's hold-out dataset. The aim of these figures is to depict the relationship between the number of organism-specific peptides in the training dataset and the corresponding mean model performance, as indicated by various evaluation measures. The number of organism-specific peptides in the training dataset are plotted against the estimated mean performance according to different performance indicators.

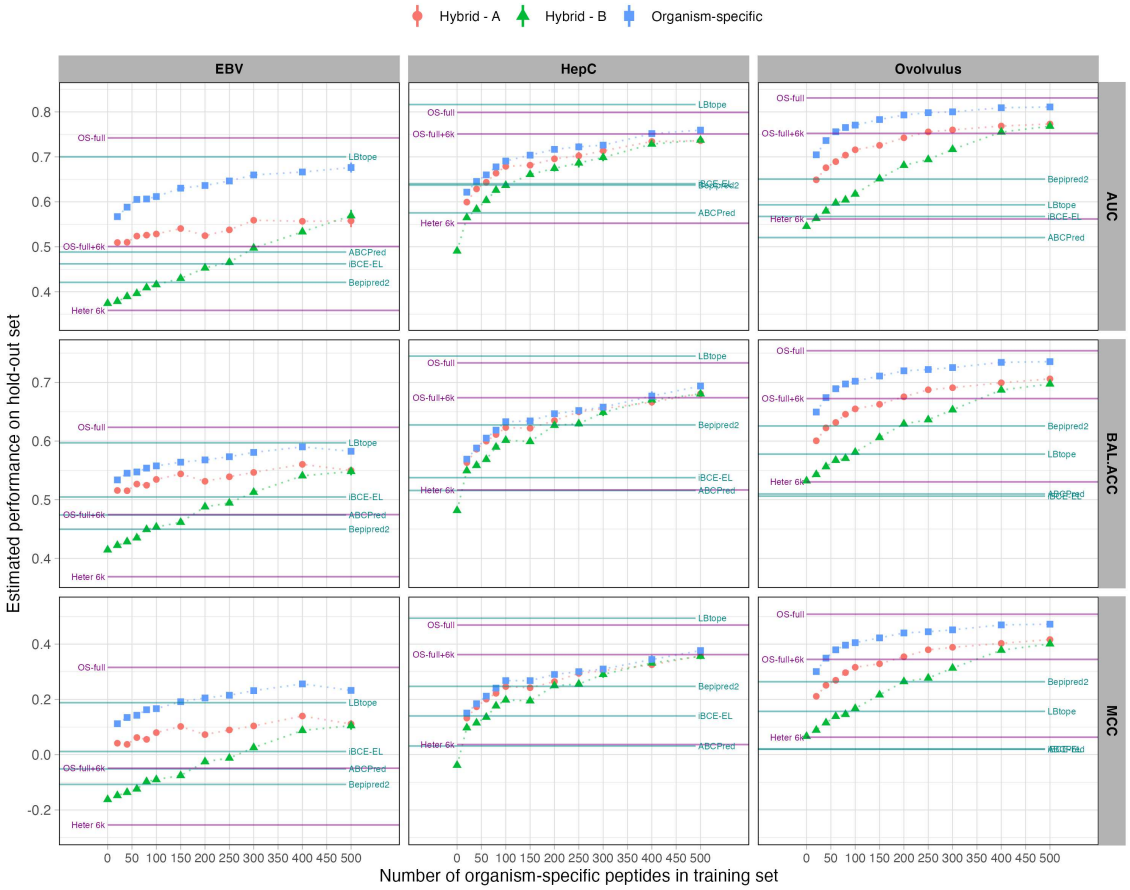


Figure 20. Mean performance scores (Area Under the Curve (AUC), balanced accuracy (BAL_ACC), Matthew's Correlation Coefficient (MCC)) and their corresponding standard errors for all models tested on the organism-specific hold-out datasets. Blue squares indicate the scores from the *organism-specific* models (trained on organism specific datasets). Red circles denote scores from the models trained on the *Hybrid A* (50 % org-spec, 50 % other/doubled data) datasets. Green triangles indicate scores from the models trained on the *Hybrid B* (1000 peptide) datasets. Horizontal lines depict reference value scores for each organism including from models trained on the complete organism-specific dataset (*'OS-full'*); on a large heterogeneous dataset (*'Heter 6K'*); on a large hybrid dataset (*'OS-full+6K'*); and scores from several predictors from the literature on the same hold-out sets. (Note: Standard error bars are often shorter than the size of the marker representing the point estimate.)

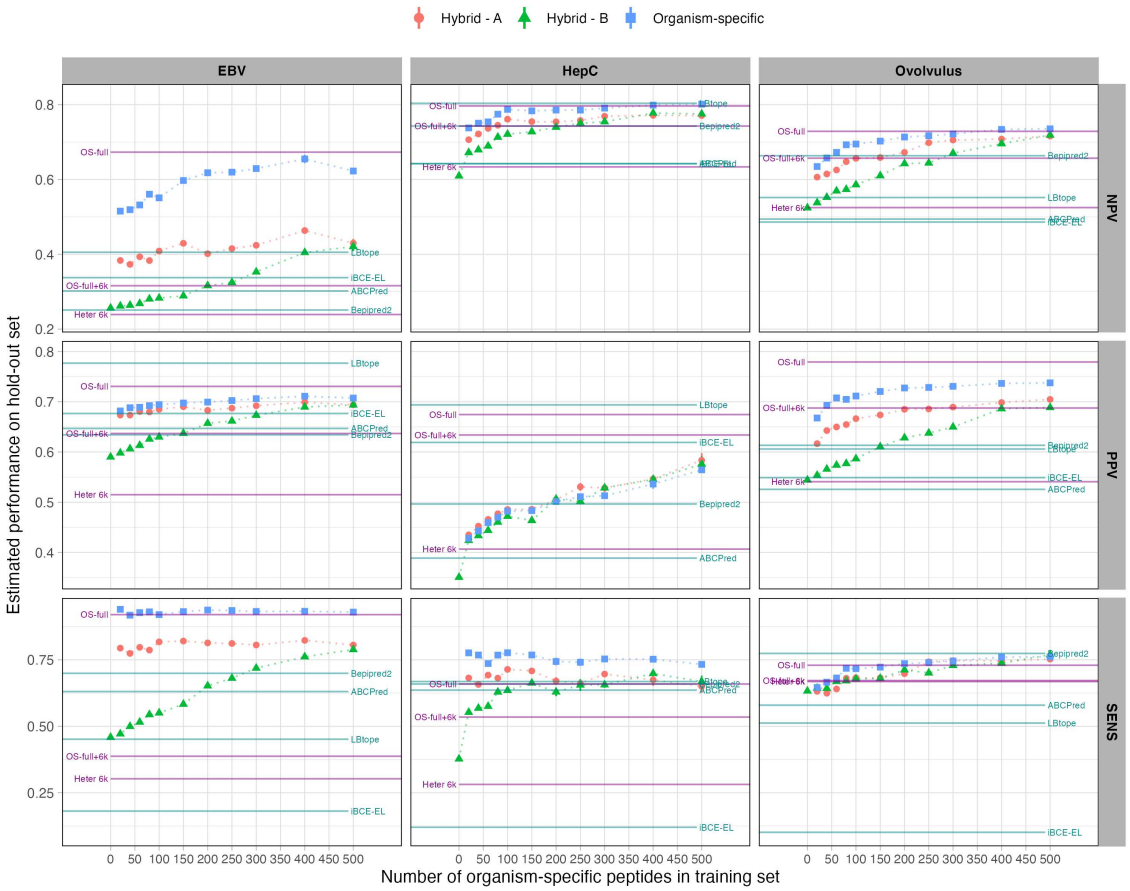


Figure 21. (Continuing from Figure 20) Mean performance scores (Positive Predictive Value (PPV), Negative Predictive Value (NPV), Sensitivity (SENS)) and standard errors for all models tested.

Results from figures 20 and 21 clearly show that, for all pathogens tested, organism-specific training resulted in uniformly better performance across all dataset sizes when compared to models trained on hybrid or purely heterogeneous data, even when as few as 20 organism-specific peptides are used in the training set. Additionally, the performance of organism-specific models swiftly surpasses that of most comparison predictors, even when only a limited number of organism-specific peptides are available for model training. The strong performance of LBtope on the Hepatitis C data, as evident across all performance indicators, can be partly explained by the overlap between some of the hold-out peptides used in this study and LBtope's training data [57] (As detailed in Section 4.8.2).

Looking in more detail at the *Onchocerca volvulus* results (the largest dataset in this study), the highest scores on the hold-out set are achieved by the full organism-specific model (*OS-full*) for all performance indicators except for sensitivity. Following this, the split-sampling organism-specific models (*organism-specific*), generally secure the next highest scores, converging towards the performance of the full organism-specific model after approximately 150 peptides. A clear pattern can be seen across all performance

measures: models trained on the organism-specific datasets consistently and uniformly outperform those trained on *Hybrid-A* (double size) datasets, which in turn outperform the models trained on *Hybrid-B* (1000-peptides) datasets. The largest datasets from both *Hybrid-A* and *Hybrid-B* consistently yield very similar scores, as anticipated. This is due to the fact that in both cases, the sets consist of 500 peptides sourced from the target pathogen data and 500 peptides from the non-target pathogen data.

Within each group of tested models (*Organism-specific*, *Hybrid-A* and *Hybrid-B*) the pattern of performance improvement as the training set becomes larger was observed, as expected. Remarkably, most small-sample organism-specific models surpass those trained on the extensive heterogeneous (*Heter 6k*) and large hybrid (*OS-full+6k*) models (for ≥ 40 peptides). Additionally, for most performance indicators, the *Onchocerca volvulus* organism-specific models outperform most benchmark models from the literature, with the exception of sensitivity where Bepipred2.0 outperforms others, and NPV where Bepipred2.0 prevails over models trained with ≤ 40 organism-specific peptides.

A similar trend is evident for the models trained on the Epstein-Barr Virus data. Figures 20 and 21 once again illustrate that the full organism-specific model (*OS-full*) consistently achieves the highest scores on the hold-out set across all performance indicators, except for positive predictive value (PPV) where LBtope attains the highest score, and sensitivity (SENS) where the full organism-specific model (*OS-full*) and the reduced data split-sampling organism-specific models (*Organism-specific*) have comparable scores across all training data sizes. The overall pattern observed for the EBV models mirrors that of the *Onchocerca volvulus* models: organism-specific $>$ *Hybrid-A* $>$ *Hybrid-B*. Additionally, as the training data become scarcer, the model performance generally decreases, as anticipated. Notably, the Epstein-Barr Virus small-sample organism-specific models (*Organism-specific*) outperform all tested models from the literature (except for LBtope), as well as the *Heter 6K* and *OS-full+6K* models.

Finally, the results from the models trained on the Hepatitis C Virus data reinforce the recurring performance trends observed for the other pathogens. The LBtope predictor demonstrates exceptional performance here, achieving the highest performance scores for 5 out of the 6 selected performance measures. However, as documented in Section 4.8.2, the notably strong performance of LBtope for this pathogen, across all performance measures, can be partially attributed to the presence of several hold-out peptides in LBtope's training data[1]. With the exception of LBtope's results, the pattern observed for the Hepatitis C models closely resembles the results seen in the other two pathogens. Organism-specific

[1] `https://webs.iiitd.edu.in/raghava/lbtope/data/LBtope_Variable_Positive_epitopes.txt`

models tend to outperform the literature predictors even when trained with a relatively modest number of peptides, typically between 40 and 100, depending on the performance measure. For the Hepatitis C Virus epitope prediction models, the performance discrepancy within each group is notably smaller than the variations observed for the other tested organisms, although there remains a clear trend of organism-specific > *Hybrid-A* > *Hybrid-B* across all performance indicators except for PPV, where the three training approaches generally overlap across all data sizes.

When comparing the scores of all organism-specific reduced-data models (*Organism-specific*) to the scores of the purely heterogeneous models (*Heter 6K* & left-most point in the *Hybrid-B* group) across all organisms and performance measures, Figures 20-21 clearly demonstrate that almost all organism-specific models achieve significantly higher scores. Additionally, when comparing the organism-specific models to the hybrid models (*Hybrid-A*, *Hybrid-B* and *OS-full+6K*) and to the generalist predictors from the literature, generally the organism-specific models exhibit superior performance. This performance advantage is evident even when the organism-specific models are trained on relatively modest-sized datasets. Across all cases, the prediction performance diminishes as the number of organism-specific peptides in the training set decreases, even when the total number of peptides in the training set remains constant (*Hybrid-B*). Notably, the organism-specific models in this study exhibit a higher degree of robustness, displaying smaller performance declines as the amount of organism-specific data decreases, compared to *Hybrid-A* and particularly *Hybrid-B*.

5.6 Discussing the Limits of Organism-Specific Training

The findings of this study indicate that, when compared to heterogeneous and hybrid training approaches, organism-specific training yields higher performance scores for linear B-cell epitope predictors, even when training with very small dataset sizes. The impact of the number of organism-specific peptides in the training set on the predictive performance of organism-specific models is profound, especially up to around 100 or 150 peptides. Beyond this point, the performance improvement becomes less pronounced, reaching a point of diminishing returns as more data is added, ultimately approaching the levels achieved by models trained on the full organism-specific training data. These results also demonstrate that, even with limited amounts of organism-specific data, organism-specific models generally outperform generalist training models (predictors from the literature), which are trained on diverse peptides from various pathogens. The only systematic exception was the high observed performance of LBtope for the Hepatitis C Virus; where the presence of hold-out examples in LBtope's training data (9.59% of the Hep C hold-out sequences are present in the LBtope training dataset) resulted in some level of data leakage [152].

In addition to showing that organism-specific training outperforms heterogeneous and hybrid training, this work highlights that introducing unrelated data into organism-specific training sets leads to decreased generalisation performance when predicting linear B-cell epitopes for a target pathogen. Notably, as seen in Figures 20-21, the addition of heterogeneous data (as demonstrated by the *Hybrid-A* and *Hybrid-B* comparisons) consistently results in poorer prediction performance. These results highlight the potential advantages of employing organism-specific training to achieve optimal performance when developing models for organism-specific epitope prediction. These findings underscore the importance of using highly specific training data, comprising only labeled peptides from the target organism for organism-specific predictions. This approach has been demonstrated to consistently outperform models trained on heterogeneous or hybrid datasets, as well as generalist predictors from the literature. Therefore, a key takeaway is that for optimal performance in organism-specific epitope prediction, it is crucial to curate training data that is tailored to the specific organism of interest, thus maximizing the relevance and accuracy of the resulting predictive models.

In summary, the comprehensive results presented in this study strongly suggest that organism-specific models, even when trained on small datasets comprising $\geq 100 - 150$ peptides, offer highly competitive predictive performance compared to the tested generalist predictors. Additionally, the point at which organism-specific models start to outperform generalist predictors appears to vary depending on the organism. For *O. volvulus* and Epstein-Barr Virus predictors the performance of the organism-specific models generally compared favourably to that of generalist models down to the smallest organism-specific dataset tested (20 peptides); while for Hepatitis C Virus predictors, a larger number of peptides were required for the organism-specific training to become competitive. These findings not only emphasize the effectiveness of organism-specific training but also expand the applicability of the methods outlined in Chapter 4, which were limited to data-rich organisms. In contrast, this study demonstrates that organism-specific training enhances epitope prediction performance for data-poor organisms as well. For context, the number of labeled peptide examples in the full training sets used in Chapter 4 were substantial, ranging from 8,819 for *O. volvulus* to 1,702 for Hepatitis C Virus, representing some of the most data-rich organisms on the IEDB. Currently, the majority of organisms possess significantly fewer available labeled epitope examples. This study highlights the significant potential of organism-specific training to enhance prediction performance for numerous organisms with limited available data, promising a more effective approach to epitope prediction across a broader range of pathogens.

5.7 Limits of Organism-Specific Training Conclusions

Chapter 4 (Organism-Specific Modelling for Linear B-Cell Epitope Prediction) demonstrated the efficacy of organism-specific training in enhancing linear B-cell epitope prediction for organisms with abundant data. This work expands the applicability of organism-specific modeling, revealing that contrary to initial assumptions, organism-specific training remains effective even for organisms with limited data (data-poor organisms). However, this study also establishes that there are indeed limits to the scope of organism-specific training for epitope prediction. The findings outlined in this study indicate that organism-specific models trained on more than approximately ~ 150 labeled peptides tend to outperform generalist predictors trained on substantially larger, but heterogeneous, datasets. Furthermore, the study confirms the trend that predictive performance generally improves in tandem with the inclusion of more organism-specific peptides in the training data (across a wide variety of indicators). It is important to note, however, that the results presented here are based on reasonably class-balanced datasets, and the investigation did not cover models trained on highly imbalanced data. The worst case among the pathogens tested was the Epstein-Barr virus data with a 2:1 balance of classes, which does not configure extreme class imbalance. While further exploration of imbalanced classification techniques for epitope prediction could potentially broaden the scope of organism-specific training, the current results presented here, combined with the growing accessibility of computational resources, already point to a promising direction. This suggests a valuable avenue for the development of bespoke predictors for specific pathogens, even in cases involving relatively data-poor organisms, such as those related to emerging health threats or neglected pathogens.

6. Discussion

6.1 Revisiting Research Questions

Numerous research questions were explored throughout this study with the principal ones being:

i Does training linear B-cell epitope prediction models using organism-specific data lead to improved prediction performance compared to models trained on heterogeneous or hybrid data?

ii How do organism-specific models compare to well-established epitope predictors from the literature?

iii How does the quantity of available organism-specific training data impact prediction performances?

iv What is the minimum amount of organism-specific data required for organism-specific models to outperform generalist predictors?

This study demonstrated that using organism-specific data significantly enhances linear B-cell epitope prediction. It showed that organism-specific Random Forest models outperformed those trained on heterogeneous and hybrid datasets for the pathogens Epstein-Barr virus, Hepatitis C virus, and *O. volvulus*, indicating that selecting relevant data and constructing tailored models for specific pathogens is more effective than expanding and diversifying the training set. Additionally, the organism-specific models achieved performance comparable to or better than common predictors from the literature and there's the potential for even higher predictive accuracy with further model refinements and the use of more informative features. Furthermore, it has been demonstrated that organism-specific training remains effective even for organisms that have limited available data. It has been shown that models trained on organism-specific datasets, containing more than 150 labelled peptides (and in some cases, even considerably less than 150), tend to outperform generalist predictors trained on larger, heterogeneous datasets. It was also shown that model performance generally improved with the addition of more organism-specific peptides in the training data. This underscores the robustness and advantages of organism-specific training across both data-rich and relatively data-poor organisms, such as emerging health threats or neglected pathogens. This research provides a valuable approach for epitope prediction researchers dealing with data-rich and relatively data-poor organisms while recognising the continued importance of generalist predictors for pathogens for which very little or no specific data is available.

6.2 Comparison with Existing Literature

While many widely-used epitope predictors were developed under a generalist approach, capable of predicting epitopes for a wide variety of pathogens, since the beginning of this research, studies have emerged that focus on developing more specific/tailored epitope predictors. For example, in 2022 Yin *et al.,* published a study: *"A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods"* [164]. This research outlines a general framework for predicting linear B-cell epitopes that are specific to human-adapted viruses. The predictor was trained on a viral-specific dataset sourced from the IEDB by filtering on all linear viral peptides derived from B-cells of the human host. After data preprocessing, the final dataset comprised 4,975 epitope and 4,956 non-epitope instances originating from 17 different human-adapted viruses. The findings show that the predictor exhibited superior performance to state-of-the-art methods on the test set, reporting an impressive AUROC score of 0.827. This is one of the few emerging works that investigate more specific epitope prediction. The developed epitope prediction framework targets epitopes from human-adapted viruses and demonstrates superior performance in predicting epitopes from this category compared to well-established state-of-the-art predictors (Including: Bepipred2.0 [63], LBtope [57], iBCE-EL [39] & EpitopeVec [66]). Additionally, the model also showed potential at being able to reveal the viral species of the epitopes.

Another notable example is the study by Bahai *et al.,* titled *"EpitopeVec: linear epitope prediction using deep protein sequence embeddings (2021)" [66]*. In this study, it was discovered that the predictive performance of the linear B-cell epitope predictor (EpitopeVec), was influenced by the origin (viral, bacterial and eukaryotic) of the antigens. Consequently, the researchers proceeded to develop a dedicated linear B-cell epitope predictor for viral antigens. This virus-specific predictor, trained on a substantial viral dataset, demonstrated enhanced prediction performance compared to the generalist predictor. This finding again aligns with the hypothesis that tailored training, whether organism-specific or virus-specific, can lead to improved epitope prediction results. Furthermore, research conducted by Silva, Ascher, and Pires in their study titled *"Epitope1D: Accurate Taxonomy Aware B-cell Linear Epitope Prediction"* [165] also evaluated the potential benefits of taxonomy-aware training in the context of epitope prediction. This study acknowledged that work published in conjunction with this PhD thesis (*"Organism-specific training improves performance of linear B-cell epitope prediction"* [166]) served as inspiration and motivation for some of their design choices.

Researchers have also studied the frequency of application of epitope predictors, and have revealed that specific predictors are employed more frequently when predicting epitopes for particular categories of pathogens. In 2020, Raouf *et al.,* published a study entitled: *"Epitope Prediction by Novel Immunoinformatics Approach: A State-of-the-art Review"* [33]. This research investigated the frequency of the application of several well-known epitope predictors at predicting both structural and linear epitopes of B and T cells, based on the nature of the target antigen which were categorised into three groups: viral, bacterial and tumor-specific antigens. The review findings indicated that Bepipred exhibited the highest frequency for linear viral B-cell epitopes, BCpred for predicting linear bacterial B-cell epitopes and ABCpred for predicting tumor-specific B-cell epitopes [33].

The emergence of studies like these provides compelling support for the hypothesis that more specific and tailored training significantly enhances the performance of linear B-cell epitope predictors. These investigations support several key insights. They highlight the importance of data curation, specifically the selection of highly relevant data specific to a particular pathogen or group of pathogens; these tailored data approaches have led to improvements in linear B-cell epitope prediction performance. These studies have also demonstrated that more specific predictors can outperform well-established generalist (state-of-the-art) epitope prediction models when it comes to predicting epitopes for a specific target pathogen or group of pathogens. This highlights the advantages of models fine-tuned to the unique characteristics of the target pathogen or pathogen categories, resulting in more accurate predictions. Additionally, it has been shown that these specialised models may possess the ability to discern the species origin of the epitopes, suggesting the presence of distinct patterns in epitope data related to their originating species. The experiments and research detailed in this thesis align with the existing literature to further confirm and reinforce the promise of organism-specific training as a viable approach for improving the performance of epitope prediction. Collectively, these findings lend strong support to the notion that, for linear B-cell epitope prediction, a tailored approach—whether focusing on a specific organism or a category of organisms—holds the potential for significantly enhanced predictive performance. This further supports the hypothesis that organism-specific or more specific/tailored training represents an optimal strategy for epitope prediction. Such an approach enables models to capture the unique patterns and characteristics that are highly relevant to the specific context of interest, resulting in improved predictive capabilities.

6.3 Limitations and Future Work

While this thesis aims to make a significant contribution to the field of linear B-cell epitope prediction, it is important to acknowledge several limitations of this work.

A common challenge in this field is the availability of high-quality labeled epitope data. More specifically, for this work, large volumes of high-quality data related to unique pathogens. Few carefully curated epitope databases exist that offer comprehensive and precise labelled epitope data. Even widely used epitope databases such as the Immune Epitope Database (IEDB) contain relatively limited epitope data. This issue becomes even more pronounced when dealing with pathogen-specific datasets, which often also suffer from imbalanced distribution between epitopes and non-epitopes. These limitations can impact the development and evaluation of epitope prediction models, especially for organism-specific models which require specific data for training. Training on small datasets can result in model overfitting, where models may perform well on the training data but struggle to generalise and provide accurate predictions for unseen data. Additionally, for numerous pathogens, although some data may be available, there's often a large data imbalance in the data. This can introduce biases and inaccuracies model performance. Given these limitations, organism-specific training, as investigated in this work, may not yet be a viable option for organisms lacking a minimum of 150 labeled peptides with a reasonably balanced class distribution.

Addressing the challenge of data availability is crucial for improving the accuracy and applicability of organism-specific epitope prediction models, especially for emerging and neglected pathogens. To address this limitation, future research could investigate the application of transfer learning techniques. These techniques have proven to be an effective strategy when dealing with limited data by leveraging pre-trained models on larger, related datasets and fine tuning them on smaller, pathogen-specific data, potentially improving prediction performance. Building on the work of this thesis, Lindeberg Faria, a researcher at the University of Brasilia in Brazil, is currently undertaking a PhD project investigating the potential of transfer learning as an approach for this challenge. In addition to transfer learning, the investigation of synthetic data generation or data augmentation could be valuable. These approaches involve artificially increasing the size of available epitope datasets, which may also help mitigate data imbalances, potentially enhancing model performance. By generating synthetic epitope data, researchers can create more comprehensive training sets for organism-specific models, ultimately improving their predictive capabilities.

Another potential limitation of this work is the generalisability of the organism-specific approach for epitope prediction. While this thesis clearly demonstrates the effectiveness of organism-specific training for specific pathogens, it may not fully investigate the applicability of these models across a broad range of organisms. Additionally, this thesis concentrates solely on the application of organism-specific modelling to predict linear B-cell epitopes and does not address the potential of this approach for conformational B-cell epitope prediction or other types of epitope prediction. Further research should aim to assess the applicability of this approach across a broader spectrum of organisms, expanding beyond the pathogens studied in this work. Extending this approach to conformational epitopes could open up new avenues for improved epitope prediction.

This thesis employs a specific curated set of features derived from the amino acid sequences of the proteins being queried. It does not delve extensively into feature engineering or consider more complex data representations, which could potentially enhance prediction accuracy. Exploring potentially more informative features including those beyond sequence-based representations may help to extract more meaningful information from the limited available data. Future research should investigate these possibilities. Incorporating new structural-based features could prove useful for organism-specific training, especially for conformational epitope prediction. Furthermore, the recent advancements in large language models for protein analysis, such as ProteinBERT [167] and ESM-2 [168, 169], offer exciting prospects for potentially improving epitope prediction. These models leverage the power of deep learning to capture intricate relationships within protein sequences and structures. Integrating such features into models for epitope prediction could potentially yield more accurate and biologically relevant results. Future research should focus on feature engineering and consider cutting-edge data representations, including structural and language model-based features, to advance the field of epitope prediction. These approaches have the potential to enhance prediction accuracy and provide deeper insights into epitope prediction. Additionally, this thesis employed Random Forest models as predictors, which are relatively simple machine learning models. Future research endeavors should explore a broader range of modeling approaches. This includes investigating the potential advantages of employing more complex models, such as deep-learning models, to enhance predictive performance. Evaluating the benefits and trade-offs associated with these alternative modeling techniques should be a focus for further investigations in the field of epitope prediction.

This thesis primarily focuses on the computational prediction of linear B-cell epitopes, it does not delve into the critical phase of clinical/experimental validation of these predicted epitopes, which, in the realm of vaccine and drug development, is an indispensable step. Clinical validation involves rigorous laboratory experiments and testing to confirm whether the predicted epitopes are indeed antigenic and trigger an immune response. Computational predictions, while valuable for initial screening and prioritisation, are not sufficient on their own to guarantee the effectiveness of a vaccine or drug candidate, making this validation process vital. Following the completion of this thesis, subsequent research has successfully applied more specific epitope predictors in the context of the Monkeypox virus in the work: *"Phylogeny-aware linear B-cell epitope predictor detects candidate targets for specific immune responses to Monkeypox virus"* [170]. This work aimed to identify candidate epitopes for the Monkeypox virus. The research uncovered nine potential peptides specific to the Monkeypox virus, among which one was a previously known and experimentally validated diagnostic target for Monkeypox, highlighting the effectiveness of more tailored epitope prediction models in specific pathogen contexts. Further experimental validation, conducted in collaboration with research partners in Brazil, confirmed that 8 out of the 9 peptides identified were indeed immunogenic, with at least 3 resulting in specific identification of the Monkeypox virus. While computational methods have provided a valuable list of potential epitope candidates for investigation, clinical validation remains the gold standard for determining whether a predicted epitope can be translated into a successful vaccine or therapeutic agent. Clinical validation involves a series of activities, including peptide synthesis, immunological assays, animal studies, and human clinical trials. These steps are essential to assess the safety, efficacy, and immunogenicity of the predicted epitopes in real biological systems. This thesis aims to contribute to the initial stage of epitope identification by identifying potential epitope candidates for use in medical diagnostics, vaccines and immunotherapies. Nevertheless, it's important to emphasise that the experimental validation steps are indispensable for these real-world applications.

6.4 Conclusions

In conclusion, this thesis aimed to explore the potential benefits of organism-specific training for linear B-cell epitope prediction. The primary goal was to investigate whether tailoring prediction models to individual organisms could yield improved predictive performance compared to generic models trained on heterogeneous or hybrid data.

Some of the main investigations of this thesis were:

- **Organism-Specific Generalisation Performance**: Assessing the generalisation performance of organism specific models/assessing how well the models perform when predicting new epitopes within proteins belonging to the specific organisms for which they were trained.
- **Exclusive Organism-Specific Training**: Investigating the effect of using exclusively organism-specific training data by comparing the predictive performance of organism-specific models to hybrid and heterogeneous models.
- **Performance Bench-marking**: Comparing the performance of organism-specific models against conventional approaches found in the existing literature.
- **Organism-Specific Data Quantity Analysis**: Investigating how the quantity of organism-specific training data influences prediction performance.
- **Limits of Organism-Specific Training**: Assessing the minimal amount of organism-specific training data required to achieve superior model performance compared to models trained on heterogeneous and hybrid datasets.

These investigations collectively contributed to a comprehensive evaluation of the potential benefits and optimal strategies for organism-specific training for linear B-cell epitope prediction.

For multiple selected pathogens, it has been shown that organism-specific Random Forest models display good generalisation performance on unseen proteins. When compared against generalist models trained on heterogeneous and hybrid data, the organism-specific models consistently outperformed their generalist counterparts across multiple performance indicators. This was also generally true when comparing organism-specific models against most well-established linear B-cell epitope prediction methods from the existing literature. Exploring the limits of organism-specific training unveiled its effectiveness across a range of data availability, including both data-poor and data-rich organisms. This study demonstrated that organism-specific models, trained on datasets containing more than approximately 150 labeled peptides, consistently outperformed their generalist counterparts trained on substantially larger and more diverse datasets. Furthermore, this investigation

highlighted the positive correlation between predictive performance and the addition of more organism-specific data. The more organism-specific data in the training data, the more robust and accurate the predictive models became.

This work provides support and evidence for the use of organism-specific training for linear B-cell epitope prediction, even for organisms with limited available data. The study's findings contribute significantly to the field of epitope prediction by showcasing the advantages of organism-specific models/tailored training. Organism-specific epitope prediction models performed as well as, if not better than, established predictors in the literature, despite their relative simplicity and the use of basic features derived solely from amino acid sequences. While this research focused primarily on simple models for linear B-cell epitope prediction, it opens the door to future investigations. It encourages the exploration of tailored approaches for conformational epitope prediction, the consideration of potentially more informative features, and the assessment of more sophisticated machine learning models for organism-specific epitope prediction. In summary, this thesis aims to enhance the field of epitope prediction, a critical step in vaccine and drug development, by demonstrating the effectiveness of organism-specific models. It helps pave the way for improved prediction tools for the discovery of diagnostic targets and vaccine candidates, ultimately contributing to advancements in health, medical diagnostic and therapeutic research applications.

References

[1] WE Paul. *Fundamental immunology. 7th edition*. London: Lippincott Williams & Wilkins, 2012.

[2] Kenneth Murphy et al. *Janeway's immunobiology. [electronic resource]*. Garland Science/Taylor & Francis Group, 2017. ISBN: 9780429084881. URL: https://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,url,shib,uid&db=cat00594a&AN=aston.b1910948&site=eds-live&authtype=ip,shib&custid=s9815128.

[3] Pauli Leinikki et al. "Synthetic peptides as diagnostic tools in virology". In: *Advances in virus research*. Vol. 42. Elsevier, 1993, pp. 149–186.

[4] Nadine L Dudek et al. "Epitope discovery and their use in peptide based vaccines". In: *Current pharmaceutical design* 16.28 (2010), pp. 3149–3157.

[5] Gregory Beck and Gail S Habicht. "Immunity and the invertebrates". In: *Scientific American* 275.5 (1996), pp. 60–66.

[6] Harvey Lodish et al. *Molecular cell biology 4th edition*. W.H.Freeman & Co Ltd, 2000.

[7] BioRender. *Antigen Recognition by Antibodies*. 2020. URL: https://app.biorender.com/biorender-templates/t-5f4fb6cc3b02b700b74df63f-antigen-recognition-by-antibodies.

[8] Jose L Sanchez-Trincado, Marta Gomez-Perosanz, and Pedro A Reche. "Fundamentals and methods for T-and B-cell epitope prediction". In: *Journal of immunology research* 2017 (2017).

[9] JL Pellequer, E Westhof, and MHV Van Regenmortel. "[8] Predicting location of continuous epitopes in proteins from their primary structures". In: *Methods in enzymology*. Vol. 203. Elsevier, 1991, pp. 176–201.

[10] Thomas J Kindt et al. *Kuby immunology*. Macmillan, 2007.

[11] Ying-Tsang Lo et al. "Conformational epitope matching and prediction based on protein surface spiral features". In: *BMC genomics* 22.2 (2021), pp. 1–16.

[12] Pernille Haste Andersen, Morten Nielsen, and Ole Lund. "Prediction of residues in discontinuous B-cell epitopes using protein 3D structures". In: *Protein Science* 15.11 (2006), pp. 2558–2567.

[13] Marc HV Van Regenmortel. “Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity”. In: *Methods* 9.3 (1996), pp. 465–472.

[14] Lenka Potocnakova, Mangesh Bhide, and Lucia Borszekova Pulzova. “An introduction to B-cell epitope mapping and in silico epitope prediction”. In: *Journal of immunology research* 2016 (2016).

[15] CB Palatnik-de-Sousa, I da S Soares, and DS Rosa. “Editorial: epitope discovery and synthetic vaccine design. Front Immunol 9: 826”. In: *Link: https://bit. ly/3pw6XbP* (2018).

[16] Dror D Siman-Tov et al. “The use of epitope arrays in immunodiagnosis of infectious disease: hepatitis C virus, a case study”. In: *Analytical biochemistry* 432.2 (2013), pp. 63–70.

[17] Oi Wah Liew et al. “Epitope-directed monoclonal antibody production using a mixed antigen cocktail facilitates antibody characterization and validation”. In: *Communications biology* 4.1 (2021), pp. 1–17.

[18] Ulrich Reineke and Mike Schutkowski. *Epitope mapping protocols*. Vol. 1. Springer, 2009.

[19] Benjamin F Arnold et al. “Integrated serologic surveillance of population immunity and disease transmission”. In: *Emerging infectious diseases* 24.7 (2018), p. 1188.

[20] Martin Closter Jespersen et al. “Antibody specific B-cell epitope predictions: leveraging information from antibody-antigen protein complexes”. In: *Frontiers in immunology* 10 (2019), p. 298.

[21] John Sidney, Bjoern Peters, and Alessandro Sette. “Epitope prediction and identification-adaptive T cell responses in humans”. In: *Seminars in immunology*. Vol. 50. Elsevier. 2020, p. 101418.

[22] “UniProt: the universal protein knowledgebase in 2021”. In: *Nucleic acids research* 49.D1 (2021), pp. D480–D489.

[23] Dennis A Benson et al. “GenBank”. In: *Nucleic acids research* 41.D1 (2012), pp. D36–D42.

[24] David L Wheeler et al. “Database resources of the National Center for Biotechnology”. In: *Nucleic acids research* 31.1 (2003), pp. 28–33.

[25] Rolf Apweiler, Amos Bairoch, and Cathy H Wu. “Protein sequence databases”. In: *Current opinion in chemical biology* 8.1 (2004), pp. 76–80.

[26] Helen M Berman et al. “The protein data bank”. In: *Nucleic acids research* 28.1 (2000), pp. 235–242.

[27] Mihaly Varadi et al. “AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models”. In: *Nucleic acids research* 50.D1 (2022), pp. D439–D444.

[28] Randi Vita et al. “The immune epitope database (IEDB): 2018 update”. In: *Nucleic acids research* 47.D1 (2019), pp. D339–D343.

[29] “UniProt, Current release statistics”. In: *EMBL-EBI homepage* (2022). URL: https://www.ebi.ac.uk/uniprot/TrEMBLstats.

[30] *Current GenBank Release Notes*. Oct. 2022. URL: https://www.ncbi.nlm.nih.gov/genbank/release/current/.

[31] Randi Vita et al. *IEDB - terms of use*. 2023. URL: https://www.iedb.org/terms_of_use_v3.php.

[32] “IEDB: Free epitope database and prediction resource”. In: *IEDB.org: Free epitope database and prediction resource* (2022). URL: https://www.iedb.org/.

[33] Ehsan Raoufi et al. “Epitope prediction by novel immunoinformatics approach: a state-of-the-art review”. In: *International Journal of Peptide Research and Therapeutics* 26 (2020), pp. 1155–1163.

[34] Ward Fleri et al. “The immune epitope database and analysis resource in epitope discovery and synthetic vaccine design”. In: *Frontiers in immunology* 8 (2017), p. 278.

[35] Thomas P Hopp and Kenneth R Woods. “Prediction of protein antigenic determinants from amino acid sequences”. In: *Proceedings of the National Academy of Sciences* 78.6 (1981), pp. 3824–3828.

[36] AS Kolaskar and Prasad C Tongaonkar. “A semi-empirical method for prediction of antigenic determinants on protein antigens”. In: *FEBS letters* 276.1-2 (1990), pp. 172–174.

[37] Johannes Söllner and Bernd Mayer. “Machine learning approaches for prediction of linear B-cell epitopes on proteins”. In: *Journal of Molecular Recognition: An Interdisciplinary Journal* 19.3 (2006), pp. 200–208.

[38] Sudipto Saha and Gajendra Pal Singh Raghava. “Prediction of continuous B-cell epitopes in an antigen using recurrent neural network”. In: *Proteins: Structure, Function, and Bioinformatics* 65.1 (2006), pp. 40–48.

[39] Balachandran Manavalan et al. “iBCE-EL: a new ensemble learning framework for improved linear B-cell epitope prediction”. In: *Frontiers in immunology* 9 (2018), p. 1695.

[40] Ying-Tsang Lo et al. “Prediction of conformational epitopes with the use of a knowledge-based energy function and geometrically related neighboring residue characteristics”. In: *BMC bioinformatics* 14.4 (2013), pp. 1–10.

[41] DJ Barlow, MS Edwards, and JM Thornton. “Continuous and discontinuous protein antigenic determinants”. In: *Nature* 322.6081 (1986), pp. 747–748.

[42] JMR Parker, D Guo, and RS Hodges. “New hydrophilicity scale derived from high-performance liquid chromatography peptide retention data: correlation of predicted surface residues with antigenicity and X-ray-derived accessible sites”. In: *Biochemistry* 25.19 (1986), pp. 5425–5432.

[43] JL Pellequer and E Westhof. “PREDITOP: a program for antigenicity prediction”. In: *Journal of molecular graphics* 11.3 (1993), pp. 204–210.

[44] Alain JP Alix. “Predictive estimation of protein linear epitopes by using the program PEOPLE”. In: *Vaccine* 18.3-4 (1999), pp. 311–314.

[45] Michael Odorico and Jean-Luc Pellequer. “BEPITOPE: predicting the location of continuous epitopes and patterns in proteins”. In: *Journal of Molecular Recognition* 16.1 (2003), pp. 20–22.

[46] Sudipto Saha and Gajendra Pal Singh Raghava. “BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties”. In: *Artificial Immune Systems: Third International Conference, ICARIS 2004, Catania, Sicily, Italy, September 13-16, 2004. Proceedings 3*. Springer. 2004, pp. 197–204.

[47] Jun Chen et al. “Prediction of linear B-cell epitopes using amino acid pair antigenicity scale”. In: *Amino acids* 33.3 (2007), pp. 423–428.

[48] Yasser EL-Manzalawy, Drena Dobbs, and Vasant Honavar. “Predicting linear B-cell epitopes using string kernels”. In: *Journal of Molecular Recognition: An Interdisciplinary Journal* 21.4 (2008), pp. 243–255.

[49] Hao-Teng Chang, Chih-Hong Liu, and Tun-Wen Pai. “Estimation and extraction of B-cell linear epitopes predicted by mathematical morphology approaches”. In: *Journal of Molecular Recognition* 21.6 (Nov. 2008), pp. 431–441. DOI: 10.1002/jmr.910. URL: https://doi.org/10.1002/jmr.910.

[50] Nimrod D. Rubinstein, Itay Mayrose, and Tal Pupko. “A machine-learning approach for predicting B-cell epitopes”. In: *Molecular Immunology* 46.5 (2009), pp. 840–847.

[51] Michael J Sweredoski and Pierre Baldi. “COBEpro: a novel system for predicting continuous B-cell epitopes”. In: *Protein Engineering, Design & Selection* 22.3 (2009), pp. 113–120.

[52] Lawrence JK Wee et al. “SVM-based prediction of linear B-cell epitopes using Bayes Feature Extraction”. In: *BMC genomics*. Vol. 11. Springer. 2010, S21.

[53] Hsin-Wei Wang et al. “Prediction of B-cell Linear Epitopes with a Combination of Support Vector Machine Classification and Amino Acid Propensity Identification”. In: *Journal of Biomedicine and Biotechnology* 2011 (2011), pp. 1–12. DOI: 10.1155/2011/432830. URL: https://doi.org/10.1155/2011/432830.

[54] Yulong Wang et al. “Determinants of antigenicity and specificity in immune response for protein sequences”. In: *BMC Bioinformatics* 12 (2011), p. 251.

[55] Bo Yao et al. “SVMTriP: A Method to Predict Antigenic Epitopes Using Support Vector Machine to Integrate Tri-Peptide Similarity and Propensity”. In: *PLoS ONE* 7.9 (Sept. 2012). Ed. by Aleksey Porollo, e45152. DOI: 10.1371/journal.pone.0045152. URL: https://doi.org/10.1371/journal.pone.0045152.

[56] Jianzhao Gao et al. “BEST: Improved Prediction of B-Cell Epitopes from Antigen Sequences”. In: *Plos One* 7.6 (2012), e40104.

[57] Harinder Singh, Hifzur Rahman Ansari, and Gajendra PS Raghava. “Improved method for linear B-cell epitope prediction using antigen’s primary sequence”. In: *PloS one* 8.5 (2013).

[58] Scott Yi-Heng Lin, Cheng-Wei Cheng, and Emily Chia-Yu Su. “Prediction of B-cell epitopes using evolutionary information and propensity scales”. In: *BMC bioinformatics*. Vol. 14. Springer. 2013, S10.

[59] Yao Lian, Meng Ge, and Xian-Ming Pan. “EPMLR: sequence-based linear B-cell epitope prediction method using multiple linear regression”. In: *BMC Bioinformatics* 15.1 (Dec. 2014). DOI: 10.1186/s12859-014-0414-y. URL: https://doi.org/10.1186/s12859-014-0414-y.

[60] Yao Lian et al. “An Improved Method for Predicting Linear B-cell Epitope Using Deep Maxout Networks”. In: *Biomedical and Environmental Sciences* 28.6 (2015), pp. 460–463. DOI: 10.3967/bes2015.065. URL: https://doi.org/10.3967/bes2015.065.

[61] Vijayakumar Saravanan and Namasivayam Gautham. “Harnessing computational biology for exact linear B-cell epitope prediction: a novel amino acid composition-based feature descriptor”. In: *Omics: a journal of integrative biology* 19.10 (2015), pp. 648–658.

[62] Weike Shen et al. “Predicting linear B-cell epitopes using amino acid anchoring pair composition”. In: *BioData mining* 8.1 (2015), pp. 1–12.

[63] Martin Closter Jespersen et al. “BepiPred-2.0: improving sequence-based B-cell epitope prediction using conformational epitopes”. In: *Nucleic acids research* 45.W1 (2017), W24–W29.

[64] Gene Sher, Degui Zhi, and Shaojie Zhang. “DRREP: deep ridge regressed epitope predictor”. In: *BMC genomics* 18.6 (2017), pp. 55–65.

[65] Maximilian Collatz et al. “EpiDope: a deep neural network for linear B-cell epitope prediction”. In: *Bioinformatics* (Sept. 2020). Ed. by Cowen Lenore. DOI: 10.1093/bioinformatics/btaa773. URL: https://doi.org/10.1093/bioinformatics/btaa773.

[66] Akash Bahai et al. “EpitopeVec: linear epitope prediction using deep protein sequence embeddings”. In: *Bioinformatics* 37.23 (2021), pp. 4517–4525.

[67] Joakim Nøddeskov Clifford et al. “BepiPred-3.0: Improved B-cell epitope prediction using protein language models”. In: *Protein Science* 31.12 (2022), e4497.

[68] Xingdong Yang and Xinglong Yu. “An introduction to epitope prediction methods and software”. In: *Reviews in medical virology* 19.2 (2009), pp. 77–96.

[69] Björn Forsström et al. “Dissecting Antibodies with Regards to Linear and Conformational Epitopes”. In: *PLOS ONE* 10.3 (2015). Ed. by Nicholas J Mantis.

[70] Arun Prasad Pandurangan and Tom L Blundell. “Prediction of impacts of mutations on protein structure and interactions: SDM, a statistical approach, and mCSM, using machine learning”. In: *Protein Science* 29.1 (2020), pp. 247–257.

[71] Jean-Luc Pellequer, Eric Westhof, and Marc HV Van Regenmortel. “Correlation between the location of antigenic sites and the prediction of turns in proteins”. In: *Immunology letters* 36.1 (1993), pp. 83–99.

[72] Martin J Blythe and Darren R Flower. “Benchmarking B cell epitope prediction: underperformance of existing methods”. In: *Protein Science* 14.1 (2005), pp. 246–248.

[73] Urmila Kulkarni-Kale, Shriram Bhosle, and Ashok S Kolaskar. “CEP: a conformational epitope prediction server”. In: *Nucleic acids research* 33.suppl_2 (2005), W168–W171.

[74] Jason A Greenbaum et al. “Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools”. In: *Journal of Molecular Recognition: An Interdisciplinary Journal* 20.2 (2007), pp. 75–82.

[75] Jing Ren et al. “Tertiary structure-based prediction of conformational B-cell epitopes through B factors”. In: *Bioinformatics* 30.12 (2014), pp. i264–i273.

[76] Shuai Lu et al. “A structure-based b-cell epitope prediction model through combing local and global features”. In: *Frontiers in Immunology* 13 (2022), p. 890943.

[77] Elisabeth Gasteiger et al. *Protein identification and analysis tools on the ExPASy server*. Springer, 2005.

[78] Carl Mayers et al. "Analysis of known bacterial protein vaccine antigens reveals biased physical properties and amino acid composition". In: *Comparative and functional genomics* 4.5 (2003), pp. 468–478.

[79] Shuichi Kawashima and Minoru Kanehisa. "AAindex: amino acid index database". In: *Nucleic acids research* 28.1 (2000), pp. 374–374.

[80] Shuichi Kawashima et al. "AAindex: amino acid index database, progress report 2008". In: *Nucleic acids research* 36.suppl_1 (2007), pp. D202–D205.

[81] Deepak Sarda et al. "pSLIP: SVM based protein subcellular localization prediction using multiple physicochemical properties". In: *Bmc Bioinformatics* 6.1 (2005), pp. 1–12.

[82] Chun-Wei Tung and Shinn-Ying Ho. "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties". In: *Bioinformatics* 23.8 (2007), pp. 942–949.

[83] Vladimir Cherkassky and Nikolaos Vassilas. "Performance of back propagation networks for associative database retrieval". In: *Int. J. Comput. Neural Net* (1989).

[84] Jason Tsong-Li Wang et al. "New techniques for extracting features from protein sequences". In: *IBM Systems Journal* 40.2 (2001), pp. 426–441.

[85] Juwen Shen et al. "Predicting protein–protein interactions based only on sequences information". In: *Proceedings of the National Academy of Sciences* 104.11 (2007), pp. 4337–4341.

[86] Jun Wang et al. "Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences". In: *International Journal of Molecular Sciences* 18.11 (2017), p. 2373.

[87] Hongchu Wang and Pengfei Wu. "Prediction of RNA-protein interactions using conjoint triad feature and chaos game representation". In: *Bioengineered* 9.1 (2018), pp. 242–251.

[88] Hongchu Wang and Xuehai Hu. "Accurate prediction of nuclear receptors with conjoint triad feature". In: *BMC bioinformatics* 16.1 (2015), pp. 1–13.

[89] G. V. Trunk. "A Problem of Dimensionality: A Simple Example". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-1.3 (July 1979), pp. 306–307. DOI: 10.1109/tpami.1979.4766926. URL: https://doi.org/10.1109/tpami.1979.4766926.

[90] Richard Bellman. "Dynamic programming". In: *Science* 153.3731 (1966), pp. 34–37.

[91] Luis O Jimenez and David A Landgrebe. "Supervised classification in high-dimensional space: geometrical, statistical, and asymptotical properties of multivariate data". In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 28.1 (1998), pp. 39–54.

[92] Laurens Van Der Maaten, Eric Postma, and Jaap Van den Herik. "Dimensionality reduction: a comparative". In: *J Mach Learn Res* 10.66-71 (2009), p. 13.

[93] Karl Pearson. "On lines and planes of closest fit to systems of points in space". In: *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2.11 (1901), pp. 559–572.

[94] Ian H. Witten et al. *Data Mining: Practical Machine Learning Tools and Techniques*. 4th ed. Amsterdam: Morgan Kaufmann, 2017. ISBN: 978-0-12-804291-5. URL: http://www.sciencedirect.com/science/book/9780128042915.

[95] Markus Ringnér. "What is principal component analysis?" In: *Nature biotechnology* 26.3 (2008), pp. 303–304.

[96] Jake Lever, Martin Krzywinski, and Naomi Altman. *Points of significance: Principal component analysis*. 2017.

[97] Ling-Yun Liu, Hong-Guang Yang, and Bin Cheng. "Prediction of Linear B-cell Epitopes Based on PCA and RNN Network". In: *2019 IEEE 7th International Conference on Bioinformatics and Computational Biology (ICBCB)*. IEEE. 2019, pp. 39–43.

[98] Bernhard Schölkopf, Alexander Smola, and Klaus-Robert Müller. "Kernel principal component analysis". In: *International conference on artificial neural networks*. Springer. 1997, pp. 583–588.

[99] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[100] Robert Hecht-Nielsen. "Replicator neural networks for universal optimal source coding". In: *Science* 269.5232 (1995), pp. 1860–1863.

[101] Nandakishore Kambhatla and Todd K Leen. "Dimension reduction by local principal component analysis". In: *Neural computation* 9.7 (1997), pp. 1493–1516.

[102] Geoffrey E Hinton and Ruslan R Salakhutdinov. "Reducing the dimensionality of data with neural networks". In: *science* 313.5786 (2006), pp. 504–507.

[103] Spencer A Thomas et al. "Dimensionality reduction of mass spectrometry imaging data using autoencoders". In: *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE. 2016, pp. 1–7.

[104] Mayu Sakurada and Takehisa Yairi. “Anomaly detection using autoencoders with nonlinear dimensionality reduction”. In: *Proceedings of the MLSDA 2014 2nd Workshop on Machine Learning for Sensory Data Analysis*. 2014, pp. 4–11.

[105] Yasi Wang, Hongxun Yao, and Sicheng Zhao. “Auto-encoder based dimensionality reduction”. In: *Neurocomputing* 184 (2016), pp. 232–242.

[106] Jorge R Vergara and Pablo A Estévez. “A review of feature selection methods based on mutual information”. In: *Neural computing and applications* 24.1 (2014), pp. 175–186.

[107] Hanchuan Peng, Fuhui Long, and Chris Ding. “Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy”. In: *IEEE Transactions on pattern analysis and machine intelligence* 27.8 (2005), pp. 1226–1238.

[108] Milos Radovic et al. “Minimum redundancy maximum relevance feature selection approach for temporal gene expression data”. In: *BMC bioinformatics* 18.1 (2017), pp. 1–14.

[109] David L Wheeler et al. “Database resources of the national center for biotechnology information”. In: *Nucleic acids research* 35.suppl_1 (2007), pp. D5–D12.

[110] “UniProt: the Universal Protein knowledgebase in 2023”. In: *Nucleic Acids Research* 51.D1 (2023), pp. D523–D531.

[111] Johannes Söllner. “Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins”. In: *Journal of Molecular Recognition: An Interdisciplinary Journal* 19.3 (2006), pp. 209–214.

[112] Christelle Pommié et al. “IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties”. In: *Journal of Molecular Recognition* 17.1 (2004), pp. 17–32.

[113] Claude E Shannon. “A mathematical theory of communication”. In: *The Bell system technical journal* 27.3 (1948), pp. 379–423.

[114] D Osorio, P Rondon-Villarreal, and R Torres. “Peptides: A package for data mining of antimicrobial peptides.” In: *The R Journal.* 7.1 (2015), pp. 4–14. URL: `https://rdrr.io/cran/Peptides/`.

[115] Gabriele Cruciani et al. “Peptide studies by means of principal properties of amino acids derived from MIF descriptors”. In: *Journal of Chemometrics* 18.3-4 (2004), pp. 146–155.

[116] Akinori Kidera et al. “Statistical analysis of the physical properties of the 20 naturally occurring amino acids”. In: *Journal of Protein Chemistry* 4 (1985), pp. 23–55.

[117] Maria Sandberg et al. “New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids”. In: *Journal of medicinal chemistry* 41.14 (1998), pp. 2481–2491.

[118] Guizhao Liang and Zhiliang Li. “Factor analysis scale of generalized amino acid information as the source of a new set of descriptors for elucidating the structure and activity relationships of cationic antimicrobial peptides”. In: *QSAR & Combinatorial Science* 26.6 (2007), pp. 754–763.

[119] Feifei Tian, Peng Zhou, and Zhiliang Li. “T-scale as a novel vector of topological descriptors for amino acids and its application in QSARs of peptides”. In: *Journal of molecular structure* 830.1-3 (2007), pp. 106–115.

[120] HU Mei et al. “A new set of amino acid descriptors and its application in peptide QSARs”. In: *Peptide Science: Original Research on Biomolecules* 80.6 (2005), pp. 775–786.

[121] Gerard JP van Westen et al. “Benchmarking of protein descriptor sets in proteochemometric modeling (part 1): comparative study of 13 amino acid descriptor sets”. In: *Journal of cheminformatics* 5.1 (2013), pp. 1–11.

[122] Li Yang et al. “ST-scale as a novel amino acid descriptor and its application in QSAM of peptides and analogues”. In: *Amino acids* 38 (2010), pp. 805–816.

[123] Alexander G Georgiev. “Interpretable numerical descriptors of amino acid space”. In: *Journal of Computational Biology* 16.5 (2009), pp. 703–723.

[124] Andrea Zaliani and Emanuela Gancia. “MS-WHIM scores for amino acids: a new 3D-description for peptide QSAR and QSPR studies”. In: *Journal of chemical information and computer sciences* 39.3 (1999), pp. 525–533.

[125] Zhi-Hua Zhou. *Machine learning*. Springer Nature, 2021.

[126] Geoffrey I Webb. *Overfitting*. 2010.

[127] Corinna Cortes and Vladimir Vapnik. “Support-vector networks”. In: *Machine learning* 20 (1995), pp. 273–297.

[128] Tin Kam Ho. “Random decision forests”. In: *Proceedings of 3rd international conference on document analysis and recognition* 1 (1995), pp. 278–282.

[129] Warren S McCulloch and Walter Pitts. “A logical calculus of the ideas immanent in nervous activity”. In: *The bulletin of mathematical biophysics* 5 (1943), pp. 115–133.

[130] Frank Rosenblatt. “The perceptron: a probabilistic model for information storage and organization in the brain.” In: *Psychological review* 65.6 (1958), p. 386.

[131] Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 2013.

[132] Bo Yao et al. “SVMTriP: a method to predict antigenic epitopes using support vector machine to integrate tri-peptide similarity and propensity”. In: *PloS one* 7.9 (2012), e45152.

[133] Leo Breiman. “Random forests”. In: *Machine learning* 45.1 (2001), pp. 5–32.

[134] EL-Manzalawy Yasser and Vasant Honavar. “Building classifier ensembles for B-cell epitope prediction”. In: *Immunoinformatics* (2014), pp. 285–294.

[135] Sudipto Saha, Manoj Bhasin, and Gajendra PS Raghava. “Bcipep: a database of B-cell epitopes”. In: *BMC genomics* 6.1 (2005), pp. 1–7.

[136] Jason Bell. “1.2.1 Supervised Learning”. In: *Machine Learning - Hands-On for Developers and Technical Professionals (2nd Edition)*. John Wiley & Sons, 2020. ISBN: 978-1-119-64214-5. URL: https://app.knovel.com/hotlink/khtml/id:kt012ET801/machine-learning-hands/supervised-learning.

[137] William S Noble. “What is a support vector machine?” In: *Nature biotechnology* 24.12 (2006), pp. 1565–1567.

[138] Simon Haykin. *Neural networks: a comprehensive foundation*. Prentice-Hall, Inc., 2007.

[139] Matt W Gardner and SR Dorling. “Artificial neural networks (the multilayer perceptron)—a review of applications in the atmospheric sciences”. In: *Atmospheric environment* 32.14-15 (1998), pp. 2627–2636.

[140] Yoav Freund and Robert E. Schapire. “Experiments with a New Boosting Algorithm”. In: *International Conference on Machine Learning*. 1996, pp. 148–156.

[141] Tianqi Chen and Carlos Guestrin. “XGBoost: A Scalable Tree Boosting System”. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. San Francisco, California, USA: ACM, 2016, pp. 785–794. ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. URL: http://doi.acm.org/10.1145/2939672.2939785.

[142] Ping Li. “Robust logitboost and adaptive base class (abc) logitboost”. In: *arXiv preprint arXiv:1203.3491* (2012).

[143] Joy Paul Guilford. “Psychometric methods”. In: (1954).

[144] Brian W Matthews. “Comparison of the predicted and observed secondary structure of T4 phage lysozyme”. In: *Biochimica et Biophysica Acta (BBA)-Protein Structure* 405.2 (1975), pp. 442–451.

[145] Gene Sher, Degui Zhi, and Shaojie Zhang. “DRREP: deep ridge regressed epitope predictor”. In: *BMC Genomics* 18.S6 (Oct. 2017). DOI: 10.1186/s12864-017-4024-8. URL: https://doi.org/10.1186/s12864-017-4024-8.

[146] Balachandran Manavalan et al. “iBCE-EL: A New Ensemble Learning Framework for Improved Linear B-Cell Epitope Prediction”. In: *Frontiers in Immunology* 9 (July 2018). DOI: 10.3389/fimmu.2018.01695. URL: https://doi.org/10.3389/fimmu.2018.01695.

[147] World Health Organization. *Onchocerciasis Fact Sheet*. https://www.who.int/news-room/fact-sheets/detail/onchocerciasis. 2019 (accessed July 22, 2020). URL: https://www.who.int/news-room/fact-sheets/detail/onchocerciasis.

[148] María-Gloria Basáñez et al. “River blindness: a success story under threat?” In: *PLoS Med* 3.9 (2006), e371.

[149] Mike Y Osei-Atweneboana et al. “Prevalence and intensity of Onchocerca volvulus infection and efficacy of ivermectin in endemic communities in Ghana: a two-phase epidemiological study”. In: *The Lancet* 369.9578 (2007), pp. 2021–2029.

[150] Sherif A. Rezk, Xiaohui Zhao, and Lawrence M. Weiss. “Epstein-Barr virus (EBV)–associated lymphoid proliferations, a 2018 update”. In: *Human Pathology* 79 (Sept. 2018), pp. 18–41. DOI: 10.1016/j.humpath.2018.05.020. URL: https://doi.org/10.1016/j.humpath.2018.05.020.

[151] Clodoveo Ferri. “HCV syndrome: A constellation of organ- and non-organ specific autoimmune disorders, B-cell non-Hodgkin’s lymphoma, and cancer”. In: *World Journal of Hepatology* 7.3 (2015), p. 327. DOI: 10.4254/wjh.v7.i3.327. URL: https://doi.org/10.4254/wjh.v7.i3.327.

[152] Shachar Kaufman, Saharon Rosset, and Claudia Perlich. “Leakage in data mining”. In: *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining - KDD’11*. ACM Press, 2011.

[153] Simon Haykin. *Neural Networks: A Comprehensive Foundation*. 2nd. Prentice Hall, 2004.

[154] F. Pedregosa et al. “Scikit-learn: Machine Learning in Python”. In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.

[155] Tianqi Chen et al. *XGBoost eXtreme Gradient Boosting*. Version 1.4.2. 2019. URL: %5Curl%7Bhttps://github.com/dmlc/xgboost%7D.

[156] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999.

[157] Henry Scheffe. *The analysis of variance*. Vol. 72. John Wiley & Sons, 1999.

[158] Karl Pearson. “LIII. On lines and planes of closest fit to systems of points in space”. In: *The London, Edinburgh, and Dublin philosophical magazine and journal of science* 2.11 (1901), pp. 559–572.

[159] Svante Wold, Kim Esbensen, and Paul Geladi. “Principal component analysis”. In: *Chemometrics and intelligent laboratory systems* 2.1-3 (1987), pp. 37–52.

[160] A. C. Davison and D. V. Hinkley. *Bootstrap Methods and Their Application*. USA: Cambridge University Press, 2013.

[161] Sture Holm. “A Simple Sequentially Rejective Multiple Test Procedure”. In: *Scandinavian Journal of Statistics* 6.2 (1979), pp. 65–70.

[162] Alexander G Georgiev. “Interpretable numerical descriptors of amino acid space”. In: *Journal of Computational Biology* 16.5 (2009), pp. 703–723.

[163] Laurens Van der Maaten and Geoffrey Hinton. “Visualizing data using t-SNE.” In: *Journal of machine learning research* 9.11 (2008).

[164] Rui Yin et al. “A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods”. In: *Briefings in Bioinformatics* 23.5 (2022), bbac281.

[165] Bruna Moreira da Silva, David B Ascher, and Douglas EV Pires. “epitope1D: accurate taxonomy-aware B-cell linear epitope prediction”. In: *Briefings in Bioinformatics* 24.3 (2023), bbad114.

[166] Jodie Ashford et al. “Organism-specific training improves performance of linear B-cell epitope prediction”. In: *Bioinformatics* 37.24 (2021), pp. 4826–4834.

[167] Nadav Brandes et al. “ProteinBERT: a universal deep-learning model of protein sequence and function”. In: *Bioinformatics* 38.8 (2022), pp. 2102–2110.

[168] Alexander Rives et al. “Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences”. In: *Proceedings of the National Academy of Sciences* 118.15 (2021), e2016239118.

[169] Zeming Lin et al. “Evolutionary-scale prediction of atomic-level protein structure with a language model”. In: *Science* 379.6637 (2023), pp. 1123–1130.

[170] Felipe Campelo et al. “Phylogeny-aware linear B-cell epitope predictor detects candidate targets for specific immune responses to Monkeypox virus”. In: *bioRxiv* (2022), pp. 2022–09.