Machine learning approaches to TCR binding prediction: benchmarking existing methods and developing a novel model for predicting TCR-MHC specificity

Machine learning approaches to TCR binding prediction: benchmarking existing methods and developing a novel model for predicting TCR-MHC specificity

by

Trupti Gore

A thesis submitted in fulfilment
of the requirements for the degree of
Doctor of Philosophy
of
Birkbeck, University of London
School of Natural Sciences

February 2025

I, Trupti Gore, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in this thesis.

Abstract

CD8+ T cells are key components of the adaptive immune system, playing crucial roles in targeting intracellular pathogens and in tumour surveillance. T cell immune function is mediated by their surface T cell receptors (TCRs), which bind to complexes formed between antigenic peptides and MHC molecules. It is estimated that a typical individual has around 10^8 unique CD8+ T cells.

The sequences of many millions of unique TCRs are known (e.g. from single cell sequencing experiments), but in all but a tiny fraction of cases their antigenic targets are unknown. This gap between the TCR sequence and TCR function affords a key motivation for the research presented in this thesis, which concerns two complementary computational prediction tasks.

The first task involved the evaluation of state-of-the-art deep learning tools that predict TCR binding to peptide-MHC complexes. This is known to be a challenging task, notably because TCR binding is modulated by six flexible loops. Tools were retrained in order to address the generalised TCR binding prediction task: can a tool predict whether a given TCR will bind to an antigenic peptide not present in the dataset used to train the tool? The results demonstrated that only two tools proved moderately successful at generalised prediction. A subsequent correlation analysis provides useful insights into the factors associated with prediction success or failure.

MHC molecules, encoded by HLA alleles, are highly polymorphic and the HLA types of the individuals from which TCR data is acquired is rarely known. A TCR is said to be HLA-restricted, i.e. it will commonly bind to complexes involving a single type of MHC. The second task addressed in this thesis involved designing a transformer-based deep learning tool for predicting TCR-HLA associations. The tool achieved overall AUCs of 0.68 using a standard 10-fold cross validation strategy on a curated dataset from all available sources. When the tool was trained on and tested on the data curated from different sources, it achieved an overall AUC of 0.76. In both cases the HLA alleles were present in both training and test sets.

Acknowledgements

“The journey of a thousand miles begins with one step” – Lao Tzu.

I couldn’t find a quote more apt than this to describe the journey that started on one fine evening with the first step—attending Birkbeck College’s Open Evening. Very little did I know then that my path would have a PhD as one of the destinations. Today, when I look back, I want to thank and express my gratitude to many people who always supported, encouraged, and stood by me.

First, I would like to express my gratitude to my supervisor and mentor, Prof. Adrian Shepherd, for his support, invaluable guidance, assistance, encouragement and indispensable advice. He has always helped me through all the difficulties I faced during this period. His role has been instrumental in this journey. Adrian, this would not have been possible without your kindness, time and patience.

I would also like to thank Justin Barton, from whom I learned a lot. You were always being so kind and helpful. This journey was a steep learning curve for me, but with your and Adrian's help, I could grasp the concepts and overcome the challenges. There is still a lot to learn and achieve, but I sincerely believe this foundation will help me to leap forward and learn more exciting things in this field.

I would also like to thank my secondary supervisor, Dr. Michele Mishto, and my thesis committee chair, Prof Katherine Thompson, for their guidance.

Secondly, my gratitude and huge thanks to my family—my parents (Aai and Baba), my siblings (Amit and Deepa), my grandma (Aaji), my lovely husband (Amol), and the apple of my eyes, my darling son Parth. You all always rooted for me, believed in me and ensured I did not deviate from my path.

I would also like to thank Birkbeck College for its crucial role in helping me to get on this train. I am here today because of the 6 to 9 pm I invested at Birkbeck over two years during my Masters. My sincere thanks to Birkbeck staff, Dr Irilenia Nobeli for

all the laughs and for getting me on the fitness journey by encouraging me to start ‘slow running’ that kept me fit enough to face the challenges during this PhD, to Dr Mark Williams for that simple conversation on Birkbeck’s Open Evening that intrigued me and ultimately led to my enrolment in the Master’s course, Prof Viji Draviam from QMUL with whom I did my first rotation project and learned a lot and achieved so many things in that short duration, and to Dr Jen Stiens, my fellow companion in this ‘journey of getting enlightened’ from Day 1 of our Master’s together.

I am immensely grateful to the Biotechnology and Biological Sciences Research Council (BBSRC) UKRI-funded studentship (BB/T008709/1) with the London Interdisciplinary Biosciences Consortium Doctoral Training Partnership Program (LIDo-DTP) for making this dream possible. I sincerely thank Nadine, Farhan (from LIDo), Tyler, Emma, and Dave (from Birkbeck) for helping me through various technical and admin-related issues. I also thank UCL for making the resources (library and compute) available.

Last but not least, I want to thank my extended family members and dear friends in the UK and back in India. I want to mention each of you here, but that will be another chapter. But I would like to mention two of my friends: Saphiya for showing me the hidden gem that is Birkbeck College, and Vandana for taking that slight detour at Birkbeck’s open evening which, unbeknownst to me at the time, put me on this path.

Also, thanks to and Oracle for Research Cloud grant [2903717], without which this computationally demanding work would not have been possible.

Table of Contents

List of Figures

List of Tables

Abbreviations

Bioinformatics

AUC	Area Under Curve
BLOSUM	Block Substitution Matrix
CNN	Convolutional Neural Network
DL	Deep Learning
FNN	Feed Forward Network
LSTM	Long Short-Term Memory
ML	Machine Learning
PDB	Protein Data Bank
RNN	Recurrent Neural Network
Unseen data (peptides or HLAs)	Peptides or HLAs that are present exclusively in the test set (i.e. Completely absent from the corresponding training dataset)

Immunology

APC	Antigen Presenting Cell
Antigen	Substances recognised by the adaptive immune system.
CTL	Cytotoxic T Lymphocytes
CMV	Cytomegalovirus Virus
EBV	Epstein–Barr Virus
Epitope	A small part of the partially degraded viral protein recognised by TCR.
ER	Endoplasmic Reticulum
HLA	Human Leucocyte Agent
MAIT	Mucosal-Associated Invariant T
MHC	Major Histocompatibility Complex
Peptide	A short sequence of amino acids as a result of cleaved antigen. In the context of TCR-pMHC binding, the terms peptide and epitope are used interchangeably.
TAP	Transporter associated Antigen Processing
TCR	T Cell Receptor
Tfh	T follicular helper
Tregs	Regulatory T cells

1 Introduction

The focus of this thesis is on a single type of molecular complex – sometimes known as the "immunological synapse"[1] – that combines three elements: an MHC class I molecule (a type of molecule associated with the nucleated cells of vertebrates); a short peptide (of diverse potential origin) bound in a groove on the surface of the MHC molecule; and a T cell receptor (TCR) found on the surface of a human CD8+ T cell (Figure 1.1). There are two main research topics, both involving computational prediction. The first concerns the prediction of whether a given TCR is likely to bind to a given combination of peptide and MHC (pMHC); this is known as the TCR-pMHC binding prediction task and is the subject of Chapter 3. The second research topic concerns the prediction of whether a given TCR is (primarily or exclusively) associated with a particular type of MHC molecule, the latter determined by the HLA genes that encode MHC class I; this is the TCR-HLA associations prediction task and is the subject of Chapter 4.

The aim of this chapter is to explain the context of these two prediction tasks from a biological perspective. It begins by explaining the broad immunological context (section 1.1) before moving on to a detailed consideration of aspects of T-cell immunology that are directly relevant to the tasks themselves. The computational background to this research is discussed in Chapter 2.

[1] After a short investigation, it appears the term may first have been used by William Paul and Robert Seder (Paul & Seder, 1994), rather than (as suggested by *Wikipedia*) Michael Dustin.

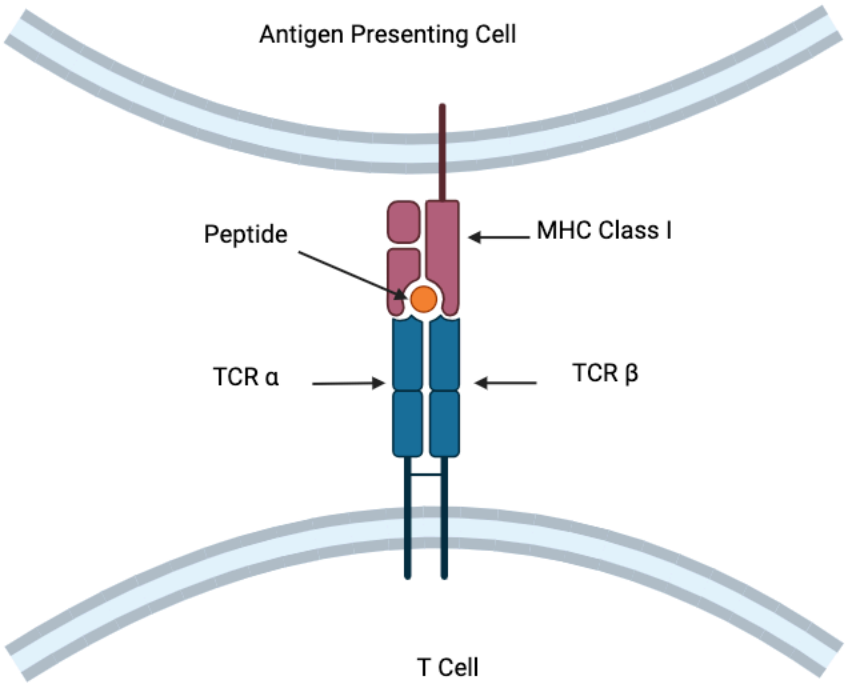

Figure 1.1 Schematic diagram of a TCR-pMHC complex. MHC class I molecules are discussed in section 1.2 and peptide-MHC binding in section 1.2.3. TCRs are discussed in section 1.3 and TCR-pMHC complexes in section 1.4. (Figure created with BioRender.com (based on figure 1 in (Yimo Sun et al., 2021))

1.1 Immunological Contexts

1.1.1 Innate and adaptive immunity

The human immune system has been described as “the most daunting example of complexity in biology” excluding the central nervous system (Paul, 2012). Broadly, it can be divided into two major systems: the innate immune system and the adaptive immune system. The former is fast but non-specific (i.e. it is not targeted at specific pathogens); it includes physical barriers such as the skin and various cells, such as macrophages, that engulf and destroy a range of pathogens, such as bacteria. The adaptive immune system, on the other hand, is slower to respond, but can target specific pathogen and prepare a rapid response against reinfection (an ability known as immune or immunological memory). The two systems communicate with each other via a “bi-directional flow of information” (McDaniel et al., 2021).

The adaptive immune system is made up of two major types of white blood cells, B cells (capable of secreting antibodies) and T cells. Some key aspects of the roles of

both B and T cells are broadly similar, namely the ability to recognise specific non-self-antigens, to contribute to the elimination of the pathogens with which those antigens are associated, and to facilitate a memory response to future infections by those pathogens. Where they differ is that B cells mainly target whole pathogens or their products (such as toxins), whereas T cells target small peptide fragments of pathogens.

There are two broad categories of T cells, and this thesis focuses on just one of them: CD8+ T cells – also known as cytotoxic T cells, cytotoxic T lymphocytes (CTLs) or killer T cells. CD8+ T cells target intracellular pathogens (such as viruses) and tumour cells and are associated with a distinct pathway of peptide presentation, the MHC class I antigen presentation pathway (see section 1.1.2). The second broad category of T cell are known as CD4+ T cells, which includes T follicular helper (Tfh) cells (which interact with B cells and facilitate the production of high-affinity antibodies) and regulatory T cells (Tregs) (which help maintain immune tolerance by suppressing the activity of other immune cells).

The “CD” in CD4+ and CD8+ stands for “cluster of differentiation” and is the prefix used in the naming of important molecules on the surface of immune cells. Both CD4+ and CD8+ T cells can be divided into subsets (two such CD4+ subsets were mentioned above, Tfh cells and Tregs), but most lie beyond the scope of this thesis (for more information about this topic and any of the wider aspects of the immune system discussed in this introductory section, see (Murphy & Weaver, 2017)).

1.1.2 The MHC class I antigen presentation pathway

Peptide fragments are presented to CD8+ T cells via the MHC class I (MHC-I) antigen presentation pathway. Within all human cells, unwanted proteins are broken down by proteasomes. Although standard cytosolic proteasomes produce peptide fragments that are presented to CD8+ T cells, the presence of proinflammatory cytokines (e.g. interferon γ) or oxidative stress induces the elevated production of immunoproteasome, a specialised proteasome that generates short peptides (capable of fitting in the MHC-I groove) with a hydrophobic or basic C-terminus (favourable for anchoring the peptide in the MHC-I groove’s terminal pocket – see section 1.2.2) (Abi Habib et al., 2022).

The cleaved peptides are then transported to the endoplasmic reticulum (ER) by TAP (Transporter associated with Antigen Processing) proteins, where a subset is loaded onto MHC-I molecules. Those that are too long may undergo trimming by an enzyme called ERAAP before loading (Murphy & Weaver, 2017). Irrespective of their length, not all peptides will be compatible with an individual's set of MHC-I molecules, as discussed below in section 1.2.3.

The peptide-MHC-I is then transported to the cell surface via the secretory pathway, where it is available to be sampled by CD8+ T cells. Note that some authors use the term T cell epitope to refer to a peptide presented in this way, whereas others reserve the term for a peptide that is involved in an active T cell response. The MHC class I antigen presentation pathway is summarised in Figure 1.2.

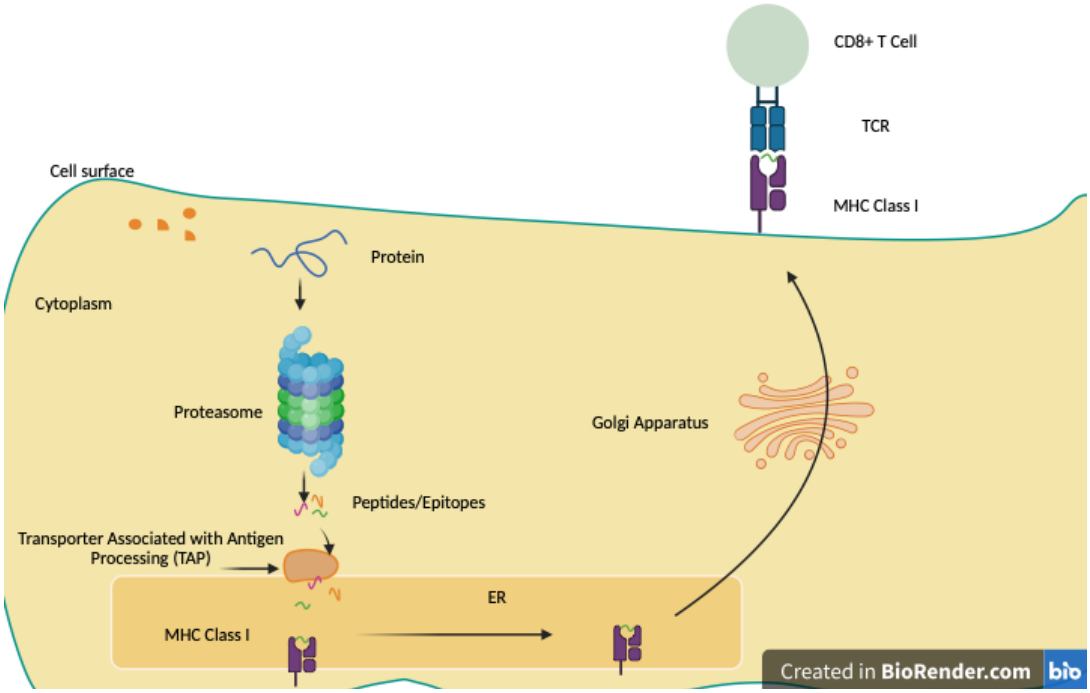

Figure 1.2 Schematic overview of the MHC class I antigen presentation pathway. Cytoplasmic proteins (both self and non-self) are cleaved into peptides by the proteasome (*upper/mid-left*). The peptides are then transported to the endoplasmic reticulum by Tapasin (TAP) and a subset is loaded onto MHC-I molecules (in the presence of chaperones) (*bottom left/centre*). The peptide-MHC complex (pMHC) is transported to the cell surface via the Golgi apparatus (*right*). The pMHC bound to the cell surface is available to be sampled by the TCRs of CD8+ T cells (*top right*). (Figure created with BioRender.com based on Figure 1 from (Yewdell et al., 2003).)

Most cells present an antigen via MHC-I molecules, whereas professional antigen-presenting cells (APCs), such as macrophages, express both MHC-I and MHC-II. A type of professional APC, called dendritic cells, has a vital role in activating CD8+ T cells. Like macrophages, dendritic cells engulf pathogens that are circulating outside cells and hence provide a mechanism for CD8+ T cells to respond to extracellular pathogens, a mechanism known as cross-presentation (Colbert et al., 2020). Cross-presentation is important for anti-tumour immunity (Wculek et al., 2020) and vaccine-induced cytotoxic immunity (Lee & Suresh, 2022). CD8+ T cell activation will be discussed further in the next section.

1.1.3 CD8+ T cell life cycle and function

Thymocytes are undifferentiated immature T cells that undergo maturation within the thymus. Maturation entails passing multiple checkpoints: the thymocyte must develop surface T cell receptors (TCRs) that are functional (β-selection), that are capable of binding to host MHCs with at least weak affinity (positive selection), and that do not bind to self-peptides with high affinity (negative selection). Failure at any of these checkpoints leads to the elimination of the cell by apoptosis. Differentiation into a CD8+ or CD4+ T cell occurs towards the end of the positive selection phase when a cell's affinity towards MHC class I or class II has been determined (Murphy & Weaver, 2017).

Only around 2% of T cells make it through all the checkpoints (Murphy & Weaver, 2017). Those that do are released into the lymphatic system or bloodstream as a naïve T cell. A naïve human CD8+ T cell may persist in the periphery for a decade or more (there is a huge variation in the estimated lifespan of both naïve and memory CD8+ T cells, as demonstrated by the contents of Table 2 in (De Boer & Perelson, 2013)).

T cell activation requires one of its TCRs to bind to an antigen in the form of a pMHC and a co-stimulatory signal, usually involving the CD28 protein expressed on the T cell surface. After activation, T cells undergo clonal expansion and differentiate into effector T cells (that immediately contribute to fighting an infection) and memory T cells (that are long-lived and provide a rapid response to reinfection via expansion and differentiation into effector T cells).

After this brief introduction to the human immune system, the MHC class I antigen presentation pathway and CD8+ T cell biology, the remainder of this chapter will focus on those aspects of MHC-I and CD8+ T cells that have a direct bearing on the research presented in chapters 3 and 4.

1.2 MHC class I molecules

1.2.1 HLA class I alleles

Human MHC molecules are encoded by HLA (Human Leukocyte Antigen) genes on chromosome 6, which are among the most polymorphic in the human genome. MHC-I molecules are divided into three major groups corresponding to three different loci: HLA-A, HLA-B and HLA-C. Currently, over 8,000 HLA-A, over 10,000 HLA-B and over 8,000 HLA-C are known (Barker et al., 2023) with many more yet to be discovered.

It is thought that most individuals are heterozygous at all three loci, with MHC heterozygosity known to be advantageous in terms of resistance to infectious disease (Penn et al., 2002) and MHC diversity implicated in human mate selection (Winternitz et al., 2017). Moreover, it is proposed that having one or more sufficiently uncommon HLA alleles is advantageous because pathogens are not under (significant) selection pressure to overcome the resistance they confer (Slade & McCallum, 1992), although the ways in which pathogens contribute to HLA allelic diversity is complicated (Spurgin & Richardson, 2010).

To gain a feel for the extent of sequence variation between HLA class I alleles, pairwise sequence alignments were performed using EMBOSS Needle (Madeira et al., 2024) using the α1/α2 domain sequences (i.e. the region that the peptide and TCR interact with) from a common HLA allele in each MHC-I group (sequences downloaded from MHC Motif Atlas at http://mhcmotifatlas.org/class1). HLA-A*02:01 and HLA-B*08:01 both have 181 residues and have 80% sequence identity; HLA-A*02:01 and HLA-C*01:02 both have 181 residues and have 82% sequence identity.

1.2.2 Overview of MHC-I structure

The MHC-I molecule is a heterodimer, with a heavy chain comprising three domains. The first two are the α1 and α2 domains, each consisting of an α-helix (around 40 residues) that forms one side of the peptide binding groove and a β-sheet (also around 40 residues) that forms part of the groove's floor. In combination, the α1 and α2 domains form a groove that (unlike the MHC-II binding groove) is closed at both ends and normally accommodates peptides that are eight to ten residues long. The α3 domain anchors the MHC molecule in the plasma membrane and interacts with the T cell's CD8 co-receptor. The light chain consists of the protein β2-microglobulin (β2m), which is positioned next to the α3 domain and contributes to MHC expression and stability (Murphy & Weaver, 2017). Ribbon diagrams of a complete MHC-I molecule plus a view of the MHC-I groove formed by the α-helices and β-sheet of the α1 and α2 domains are shown in Figure 1.3.

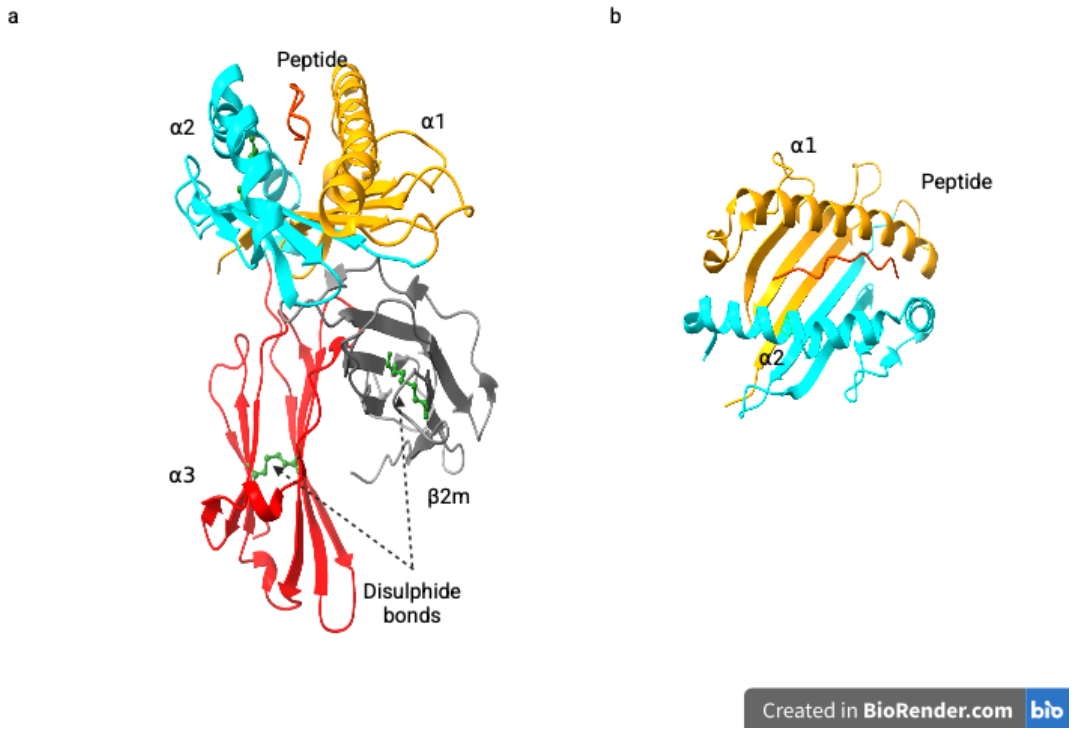

Figure 1.3 Ribbon diagrams of an example MHC-I molecule (PDB 3PWN). The peptide backbone is shown in orange. **a** Side view showing the three domains of the heavy chain (α1, α2, α3) and the light chain (β2m). **b** View from above looking down into the peptide-binding groove formed by the α1 and α2 domains. (Figure created with BioRender.com using Chimera X (Pettersen et al., 2004)

Within the MHC-I binding groove, six distinct pockets have been defined, labelled A to F, that accommodate the sidechains of the peptide binding within the groove (Saper et al., 1991). Of these, pockets B and F are considered the most important in

terms of peptide binding and the stability of the peptide-MHC complex, as they accommodate the canonical anchor residues occurring at position 2 (pocket B) and at the C-terminal end (pocket F) of the peptide (see, for example, (Harndahl et al., 2012)).

As noted in section 1.2.1, HLA class I alleles are highly polymorphic. Analysis of the location of MHC-I allelic variations performed by Murphy and Weaver (Murphy & Weaver, 2017), shows the variations mainly occur at exposed sites within the α1 and α2 domains, particularly in residues associated with the peptide-binding groove. Peptide-MHC-I binding is the topic addressed in the next section.
A recent search using the IMGT interface (Kaas et al., 2004) suggests[2] there are nearly 1,000 structures of human MHC class I molecules in the protein data bank (PDB), of which just over 30 form part of TCR-pMHC complexes, but most are not unique.

1.2.3 Peptide-MHC binding

Peptides bind within the MHC-I groove in an extended conformation. Given that the MHC-I's closed binding groove can naturally accommodate peptides up to a length of ten residues, with the two pockets that house a peptide's primary anchor residues at position 2 near the peptide's N-terminus (pocket B) and at the C-terminal end of the groove (pocket F), it appears reasonable to assume that MHC-I molecules preferentially present peptides of length eight to ten (8-mers, 9-mers and 10-mers). However, matters are rather more complex. On the one hand, elution studies have shown that 9-mers are much more common than peptides of any other length, but 10-mers and 11-mers were more common than 8-mers for four out of five common HLA alleles, the exception being HLA-B*51:01 (Trolle et al., 2016). On the other hand, large-scale binding affinity experiments carried out by the same authors suggested that HLA-B*51:01 has (on average) higher affinity for 8-mers than 9-mers and HLA-A*01:01 a higher affinity for 10-mers and 11-mers than for 9-mers. The authors explain this apparent discrepancy by arguing that, of the peptides available for binding to MHC molecules, 9-mers are more common than peptides of any other length (e.g. TAP disfavours the transportation of peptides shorter than 9 residues) (Trolle et al., 2016). Peptides longer than 10 residues are accommodated either

[2] No attempt was made to verify that all entries in the returned lists were correct.

through a reorientation of the α1 helix or by bulging out of the groove (Rudolph et al., 2006).

Irrespective of their potential preference for peptides of different lengths, differences in the pockets within the MHC-I groove determine which peptides can bind. The amino acid types that are preferred or deleterious at different positions within the binding groove are HLA allele specific. Pockets B and F that accommodate the peptide's anchor residues are generally the most restrictive. For example, nearly all peptides that bind to MHC molecules encoded by HLA-A*01:01 have a Tyrosine at position 9 (pocket F), whereas HLA-A*02:01 MHCs prefer Valine, Leucine or Isoleucine – all large hydrophobic amino acids – at the same position. The extent to which the other pockets place restrictions (if any) on the types of permitted amino acid is highly variable. This information can be summarised in an MHC binding motif (Rapin et al., 2008; Tadros et al., 2022), as shown in Figure 1.4.

Peptide-MHC binding motifs have a bearing not only on the set of peptides capable of being presented by a given allelic variant, but also on TCR recognition, as the sidechains of residues that are not buried in pockets B and F or tightly constrained in their interactions with the MHC molecule are more likely to be TCR-facing (see Figure 1.5).

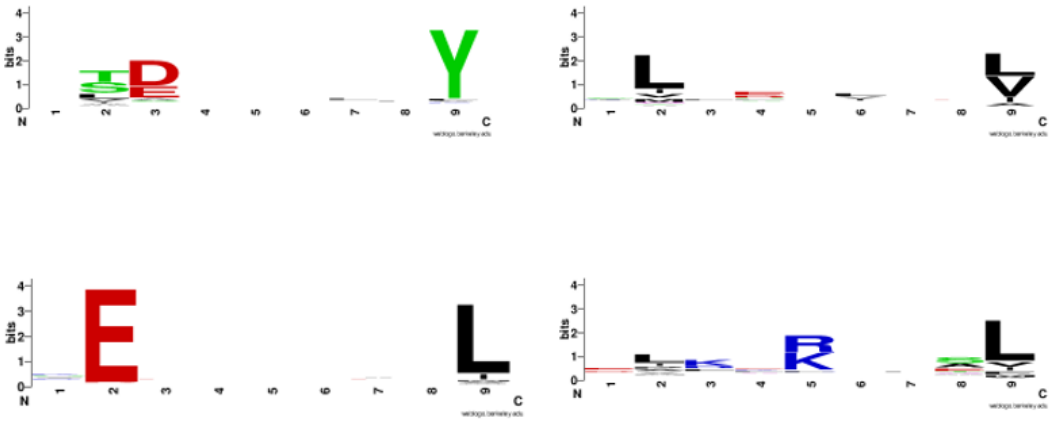

Figure 1.4 A contrasting set of MHC binding motifs. a. HLA-A*01:01: nearly all peptides have a Tyrosine at position 9, most have negatively charged amino acids (Aspartic Acid and Glutamic Acid) at position 3, and there are moderate restrictions at position 2 (mainly hydrophobic residues). b. HLA-A*02:01: most peptides have non-polar hydrophobic amino acids at positions 2 and 9 (Valine, Leucine, Isoleucine at position 9 with Methionine additionally present at position 2), with weak restrictions at positions 4 and 6. c. HLA-B*40:01: nearly all peptides have Glutamic Acid at position 2 and hydrophobic amino acids at position 9. d. HLA-B*08:01: most peptides have positively charged amino acids (Arginine, Lysine) at position 5, hydrophobic amino acids at position 9, weak/modest restrictions at positions 1, 2, 3 and 8. The peptide sequences were downloaded from MHC Motif Atlas [3] and the logos were created using WebLogo [4].

[3] http://mhcmotifatlas.org/class1

[4] https://weblogo.berkeley.edu/logo.cgi

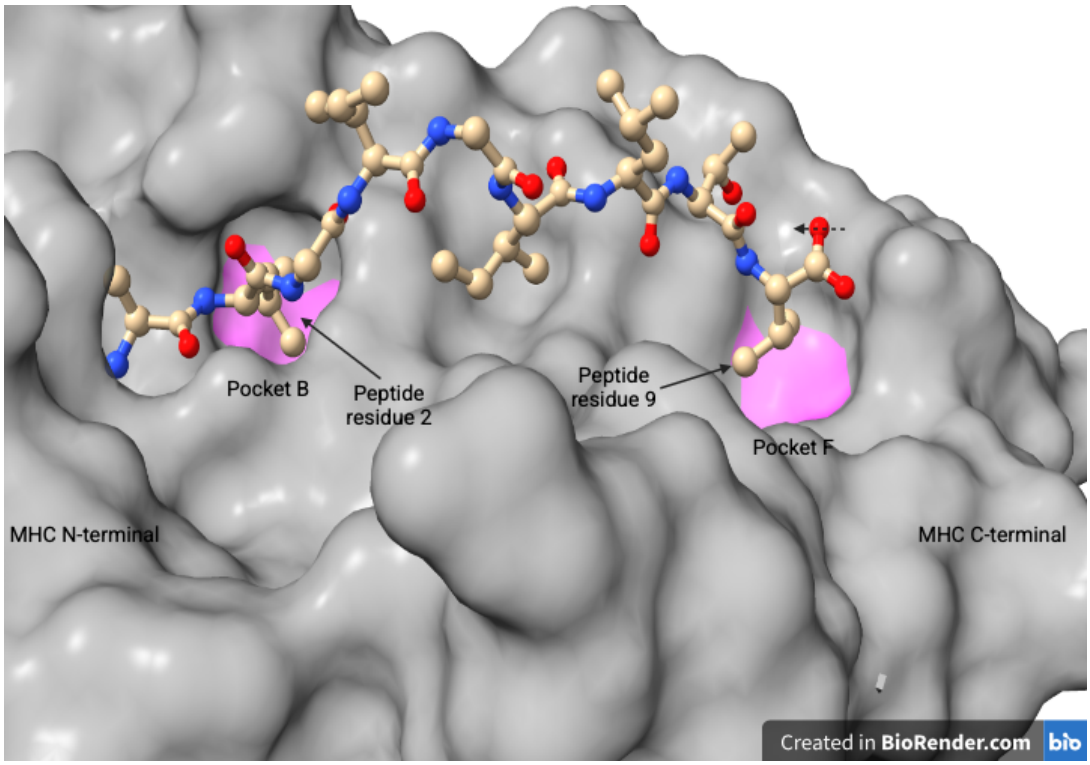

Figure 1.5 Peptide ALGIGILTV binding in the HLA-A*02:01-encoded MHC groove (PDB 1JHT). Several residues are MHC-facing, whereas the Isoleucine at position 4, Glycine at position 5, Leucine at position 7 and Threonine at position 8 are TCR-facing. . (Figure created with BioRender.com using Chimera X (Pettersen et al., 2004), based on Figure 7 from (Perez et al., 2022).)

1.3 CD8+ TCRs

1.3.1 TCR sequence diversity

A single CD8+ T cell has around 10^5 identical T cell receptors (TCRs) on its surface (Schodin et al., 1996), each formed from a pair of protein chains. In around 95% of cases, human TCRs are αβTCRs with an α chain encoded by genes in the TRA locus and a β chain encoded by genes in the TRB locus. The remainder are γδTCRs with chains encoded by the TRG and TRD loci. γδTCRs have been described as a "bridge between innate and adaptive immune responses" (Holtmeier & Kabelitz, 2005) and bind to a diverse set of ligands, rather than just peptides (Vermijlen et al., 2018); although undoubtedly interesting, they lie outside the scope of this thesis and the abbreviation "TCR" refers to an αβTCR throughout.

TCR sequences are highly diverse both within and between individuals, attributable to several factors. Each TCR chain is encoded by combining gene segments: for the α chain, this involves a variable (V), a joining (J) and a constant (C) gene segment; and for the β chain a V, a diversity (D), a J and a C gene segment. There are multiple gene segments to choose from: the TRA locus that encodes the α chain has 43-45 V, 50 J and 1 C gene segments; and the TRB locus that encodes the β chain has 40-48 V, 2 D, 12-13 J and 2 C gene segments.
(IMGT Gene Number page, accessed 19/09/2024: https://www.imgt.org/IMGTrepertoire/LocusGenes/genetable/human/geneNumber.html). In combining gene segments – a process known as V(D)J recombination – nucleotides may be added or removed at the junctions between the gene segments (Figure 1.6). Taking these factors together, it is estimated that around 10^{18} distinct TCR chains (α chains plus β chains) are theoretically possible (Murphy & Weaver, 2017). A final factor contributing to TCR diversity is the pairing of α and β chains, which has been described as “almost unconstrained” and “nearly random” (Shcherbinin et al., 2020).

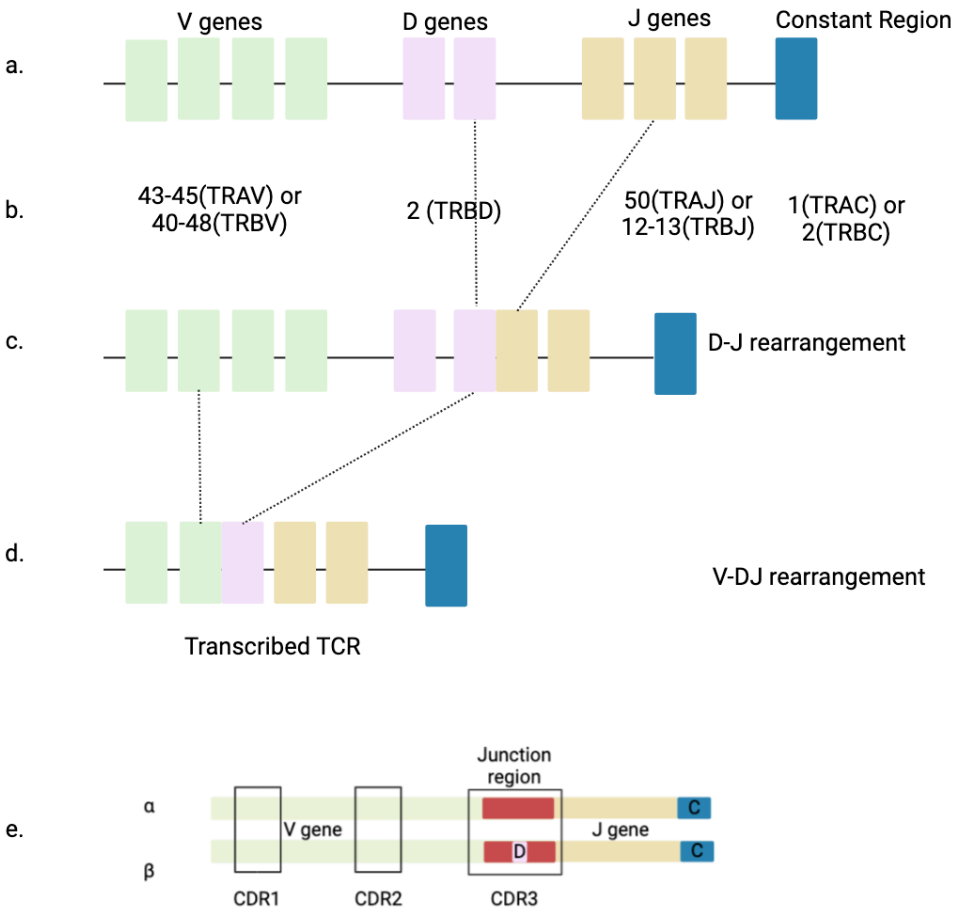

Figure 1.6 Schematic diagram of V(D)J recombination. Note that individual insertions and deletions (within the junction region) are omitted from this figure. **a.** Schematic showing the multiple germline V, D and J gene segments associated with the β chain (the α chain has no D segments). **b.** The number of germline α chain gene segments (TRAV, TRAJ and TRAC) and germline β chain gene segments (TRBV, TRBD, TRBJ and TRBC) found in humans. **c.** In forming the β chain, first, a D segment is joined with a J segment. **d.** The DJ segment from *c.* is then joined to a V gene segment and combined with a constant region. **e.** Schematic of the full TCR showing the location of the gene segments. Six hypervariable loops (CDRs), discussed below in section 1.3.2, are annotated. (Figure created with BioRender.com)

Published estimates of the number of unique TCRs (known as species richness) within a single individual vary by (at least) an order of magnitude. An early calculation concluded that there are around 2.5 x 10^7 unique TCRs consisting of 10^6 β chains paired, on average, with "at least 25 different α chains" (Arstila et al., 1999). A more recent study concluded that 20 to 35 year-olds have around 10^8 unique naïve CD8+ T cells, dropping to around 5 x 10^7 or lower in those aged 70 to 85. The diversity found in memory CD8+ T cells was around 250- to 500-fold lower (Qi et al., 2014). Given that the total number of T cells in a typical individual is estimated to be 10^{12}, it is obvious that many T cells have the same TCRs. This is mainly attributable to clonal expansion after activation (see section 1.1.3), but it is also evident that some TCR sequences are more likely to be created than others. This is clear from the higher

prevalence of public β chains in the naïve T cell repertoire than one would expect by chance, where a public sequence is usually defined as one found in more than a single individual. For example, in a study of four healthy donors with the HLA-A*02:01 allele, a quarter or more of the naïve β chain repertoire was shared with at least one of the other donors (Venturi et al., 2011). These public β chains generally occur at greater frequency within the naïve repertoires of the individuals concerned and are attributable to a process known as convergent recombination, whereby certain underlying properties (e.g. patterns of codon degeneracy) combine with differences in the efficiency with which certain recombination events occur (Venturi et al., 2011).

The extent of polymorphism within the TCR-encoding genes is poorly understood. The 23rd July 2024 IMGT reference sets (available from https://www.imgt.org/vquest/refseqh.html) contained 113 TRAV and 147 TRBV genes, but recent research involving inferences from T cell repertoire data, in which 18 TRBV alleles not recorded by IMGT were identified in a dataset of full-length sequences from 53 individuals, suggests there is extensive germline variability yet to be discovered (Omer et al., 2022).

With regards to the pattern of variability along the length of each TCR chain, the most variable sections are the junction regions that span the joins between V and J in the α chain and V, D and J in the β chain, the latter being by far the most variable region within the entire TCR. The latter corresponds with a feature of vital structural and functional importance, that will be introduced in the next section – the CDR3β.

1.3.2 Overview of TCR structure

An αβTCR is a disulphide-bonded heterodimer consisting of an α and a β chain. Both chains have the same structural components: a transmembrane region (ending in a short cytoplasmic tail) and two extracellular domains, a constant domain and a variable domain (Figure 1.7). The α and β variable domains are involved in binding to the pMHC complex and are, therefore, the key focus for the research presented in this thesis.

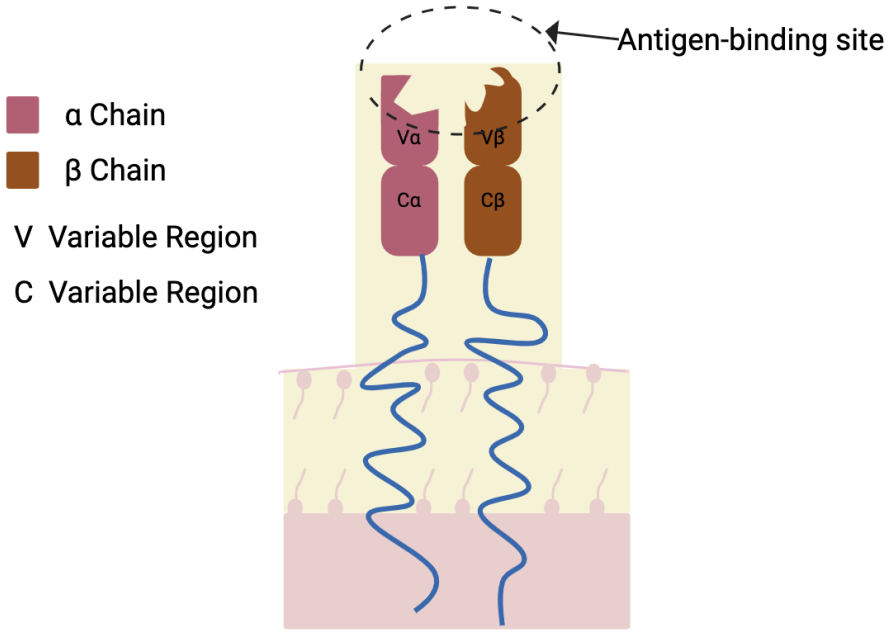

Figure 1.7 Schematic diagram showing a complete membrane-bound TCR protein.
(Figure created with BioRender.com)

A key feature of the α and β variable domains are the presence of six hypervariable loops known as CDRs (Complementarity Determining Regions) (Figure 1.8). The CDR1s and CDR2s are germline-encoded (CDR1α and CDR2α by the TRAV gene, CDR1β and CDR2β by the TRBV gene), and have sequences that are much less diverse than the CDR3s, which (as noted in section 1.3.1) span the junction regions where multiple genes segments are joined together. The CDR1s and CDR2s from both chains are comparatively short, with lengths typically in the range 5-8 residues (Wong et al., 2019), whereas both CDR3s are generally longer and more variable in length (as expected, given the occurrence of nucleotide insertions and deletions): 6-18 residues with a mean of 11.7 residues for CDR3α, and 7-20 with a mean of 12.7 for CDR3β (K. Yu et al., 2019). An additional short variable loop of 5 or 6 residues, the "non-classical" CDR2.5s located between CDR2 and CDR3 on both chains, sometimes make contact with pMHC (Dash et al., 2017).

With the notable exception of CDR3β, CDRs are known to adopt a rather limited number of distinct backbone conformations in solved structures, which has led to the definition of "canonical classes" of CDR conformation. In a 2019 analysis of over 270 high-resolution PDB structures, the number of canonical classes identified varied between one for CDR1β and seven each for CDR1α and CDR2α. However, the

situation is more complicated than this suggests. Excluded from the canonical classes are structures assigned to “pseudo-classes” containing fewer than three unique sequences, and “many cases” have been observed where different structures of the same sequence (e.g. both bound and unbound to pMHC) were assigned to different classes (Wong et al., 2019). This points to an important property of TCR CDRs, namely their flexibility. McMaster et al.'s comparisons of bound and unbound TCR loops' quantification in terms of the backbone root mean squared deviation (RMSD) show a mean change of 1.69 Å for CDR1α, 1.33 Å for CDR2α , and 2.38 Å for CDR3α. They have also reported a mean change of 0.82 Å, 0.91 Å, and 1.50Å for CDR1β, CDR2β, and CDR3β respectively (McMaster, Thorpe, Rossjohn, et al., 2024).

Although there are no CDR3β canonical classes, certain conformational constraints have been identified at the two ends of the loop (based on earlier analyses of the equivalent loop in antibodies, the heavy chain CDR3). Specifically, the so-called “torso” – consisting of the first three and last four loop residues – are observed in TCRs to adopt an “extended” conformation and rarely the “kinked” or “bulged” conformation frequently observed in antibodies (Wong et al., 2019).

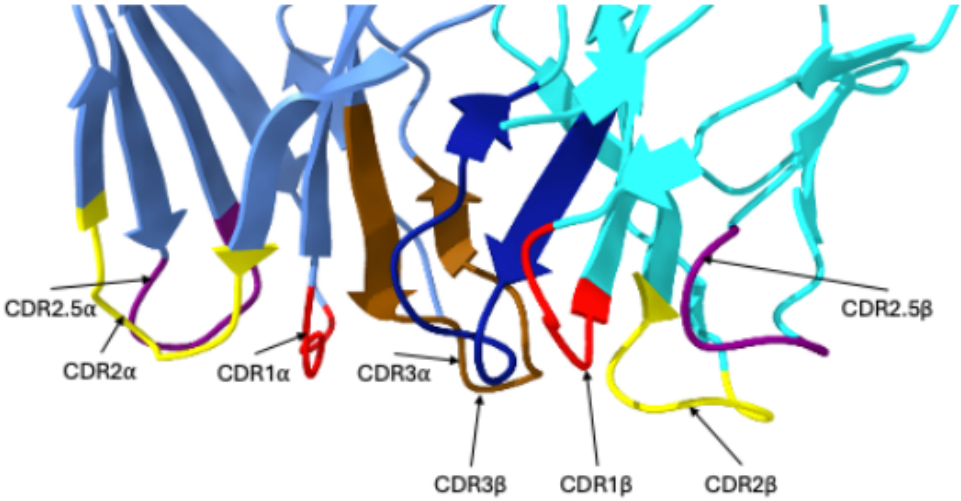

Figure 1.8 Ribbon diagram showing the TCR complementarity determining regions (CDRs), including the non-classical CDR2.5s. The CDR1s are shown in red, the CDR2s in yellow, the CDR2.5s in purple, the CDR3α in brown and CDR3β in midnight blue. The figure was recreated using Chimera X (Pettersen et al., 2004) based on Figure 4B in (Barbosa et al., 2021).

1.4 TCR-pMHC complexes

This section focuses on key aspects of how TCRs and pMHCs interact to form TCR-pMHC complexes. Whereas it is evident that many TCRs can bind to the same pMHC – indeed, more than half the TCR-pMHC samples collected for the benchmarking research presented in Chapter 3 consisted of TCRs binding to a single pMHC (involving an immunodominant EBV epitope) – it is also the case that a single TCR can bind to many pMHCs. This is the topic of section 1.4.1. The contacts formed between TCR and peptide, and between TCR and MHC are considered in section 1.4.2. Finally, the propensity of TCRs to bind to a single type of MHC molecule (known as HLA restriction) is considered in section 1.4.3.

To provide a structural context for these topics, a ribbon diagram of a CD8+ TCR-pMHC complex is shown in Figure 1.9.

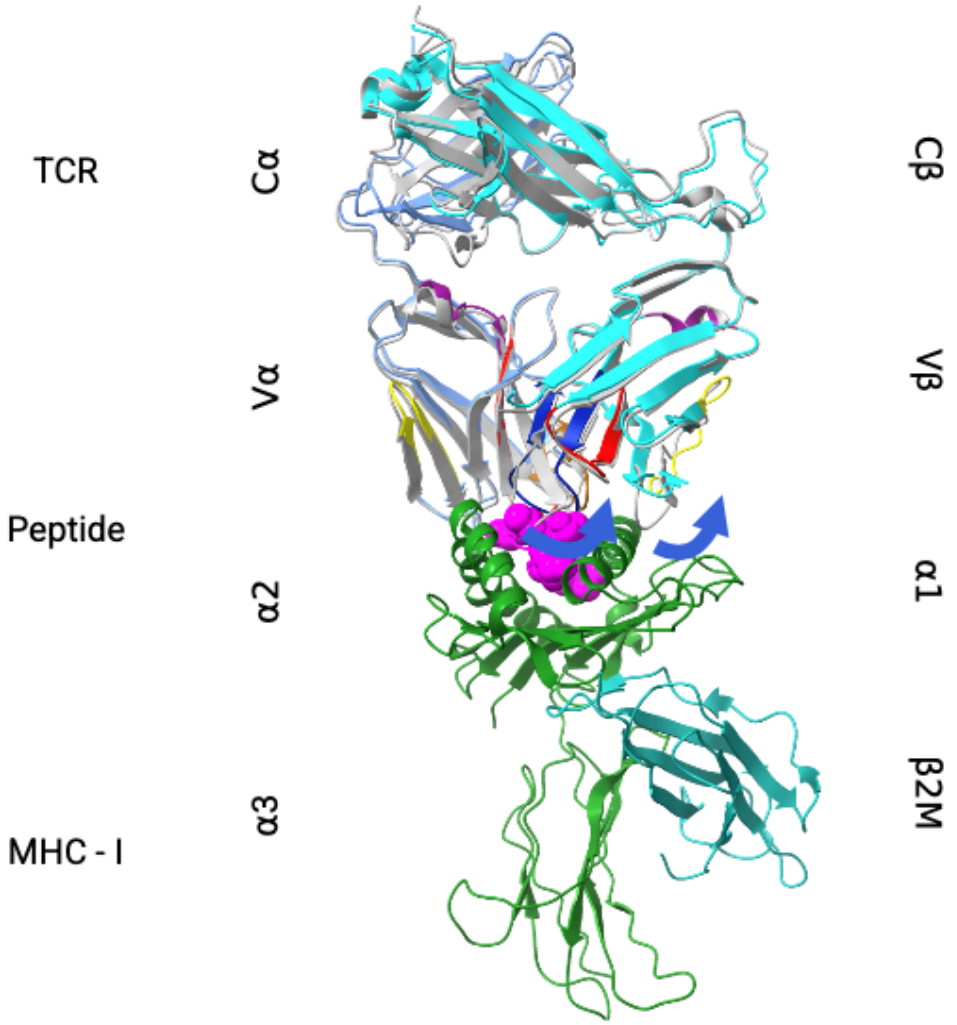

Figure 1.9 Ribbon diagram of a CD8+ TCR-pMHC complex (PDB 4JFH (unbound) superimposed with bound structure 4JFD). The major components of the MHC molecule (encoded by HLA-A*:02:01) are as follows: the heavy chain of 4JFD (α1, α2 and α3) domains are in green, and the light chain (β2-microglobulin) in cyan. Peptide ELAAIGILTV is in magenta. The TCR α chain for 4JFD is in cornflower blue and β chain in cyan. Unbound PDB complex 4JFH is coloured in grey. It shows the movement between the bound and unbound TCR structures. (The figure was recreated using Chimera X (Pettersen et al., 2004) based on Figure 2A in (McMaster, Thorpe, Rossjohn, et al., 2024).

1.4.1 TCR cross-reactivity

Given that the number of unique TCRs within a single individual is around 10^8 (as discussed in section 1.3.1) whereas "the simple arithmetic of effective immunity requiring the recognition of $>10^{15}$ potential foreign peptides" (Sewell, 2012), it has been argued that near-universal TCR cross-reactivity[5] – i.e. the ability of nearly all

[5] Other widely used terms are TCR promiscuity, TCR degeneracy and TCR polyspecificity.

TCRs to bind to many different pMHC complexes – is a biological necessity.(Mason, 1998). To quantify the potential extent of a single TCR's ability to bind multiple pMHCs, an experimental study by Wooldridge and co-workers demonstrated that a single CD8+ T cell clone was capable of binding to more than one million 10-mers bound to HLA-A*02:01-encoded MHC molecules (Wooldridge et al., 2012).

Various factors are known to contribute to TCR cross-reactivity, including: the ability of a TCR to change its position (binding register) on the MHC groove and/or its angle of binding (Baker et al., 2012); the flexibility of a TCR's CDRs ((Armstrong et al., 2008); see section 1.3.2)); the fact that most TCRs form contacts with a limited number of TCR-facing peptide sidechains (see Figure 1.6); and the flexibility of the pMHC complex (Borbulevych et al., 2009).

It is worth noting the striking contrast between certain cases of TCR cross-reactivity versus those of TCR specificity. For example, there are cases where a single TCR may bind to peptides that have limited or even no sequence overlap (S. Zhang et al., 2015), whereas in other cases, a single amino acid change at an anchoring position may be sufficient to prevent a TCR from binding. The latter is surprising, as the sidechain of an anchoring residue is generally buried in a pocket and therefore hidden from the TCR (see section 1.3.2), but Smith and co-workers found an example where a TCR failed to bind to an anchor-modified peptide because the residue change induced a change to the interactions between TCR and MHC "in distant parts of the interface" (Smith et al., 2021).

1.4.2 TCR-peptide and TCR-MHC structural contacts

Although there are some exceptional, "rule-breaking" examples of TCR-pMHC complex formation (some of which are documented in (Szeto et al., 2020)), widely applicable constraints have been observed about the way that TCRs engage with pMHC complexes.

One of the most fundamental constraints is the relative orientation of TCR and pMHC. Until comparatively recently, it appeared to be universally the case that TCRs bind on top of the MHC binding groove, adopting a diagonal orientation with respect to the groove such that the TCR α chain is located towards the N-terminal end of the peptide. A 2011 analysis of 61 TCR-pMHC structures concluded that the TCR

crossing (or docking) angle varied between 20° and 87° (J. M. Khan & Ranganathan, 2011). However, a TCR was subsequently discovered that adopts a reversed orientation ("a 180° reversed polarity") compared to all previously observed crossing angles; the wider biological significance of reverse docking remains uncertain (Gras et al., 2016). An example of a "canonical" TCR footprint superposed on a pMHC-I complex is shown in Figure 1.10.

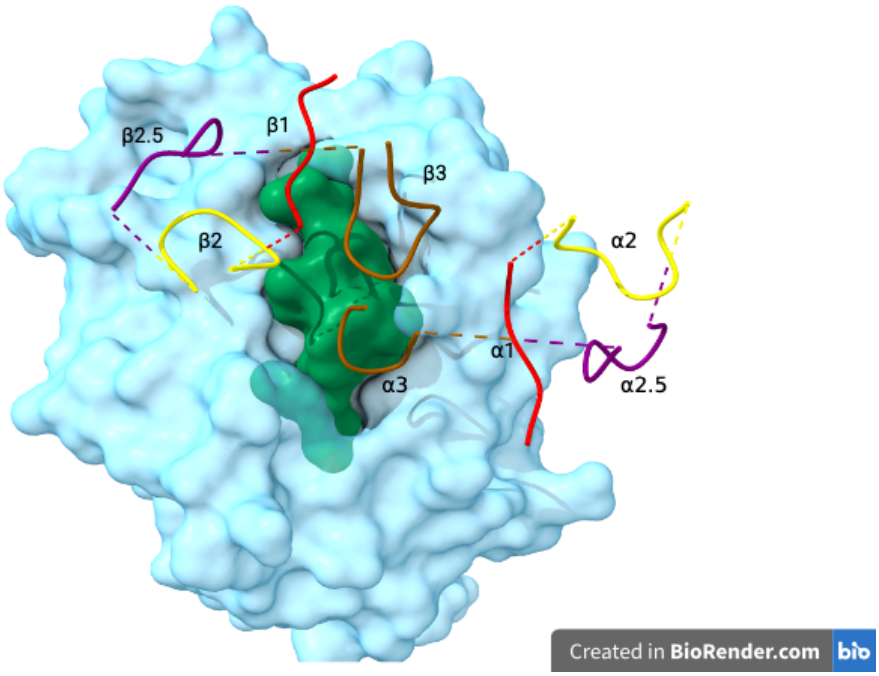

Figure 1.10 Footprint of a TCR bound to a peptide-MHC-I complex. HLA-A*02:01-encoded MHC molecule (in sky-blue) is presenting immunodominant peptide GILGFVFTL (in green), an Influenza A matrix protein (PDB 1OGA). This arguably represents the canonical orientation of the TCR with respect to the pMHC complex. The coloured areas show the positions of the TCR's CDRs, with CDR3s (brown) positioned over the peptide and the other CDRs - CDR1s (red), CDR2s (yellow) and CDR2.5s (purple) positions over the MHC. (Figure created with BioRender.com using Chimera X (Pettersen et al., 2004)

TCR contacts with the peptide are mainly via the CDR3s (the most diverse of the CDRs, as explained in section 1.3.2). Owing to the crossing angle of the TCR, both CDR3 loops are equally involved in peptide binding with the CDR3α is typically in contact with the N-terminal half of the peptide and the CDR3β with the C-terminal half of the peptide (McMaster, Thorpe, Rossjohn, et al., 2024). As the more variable and typically slightly longer CDR (see section 1.3.2), CDR3β is often assumed to make more contacts with the peptide and have the dominant role in determining TCR specificity. This is often the case, but CDR3α has been shown to play a dominant

role (and CDR3β a “minor” role) in TCRs directed against an immunodominant EBV epitope (Gil et al., 2020).

TCR contacts with MHC involve multiple CDRs. The conventional view has been that the CDR1s and CDR2s play the dominant role in MHC engagement via contacts with the exposed upper surfaces of its α-helices. Given the TCR crossing angle, CDR1α and CDR2α mainly engage with the α2-helix, whereas CDR1β and CDR2β mainly engage with the α1-helix. However, a mutagenesis study of TCRs binding to an immunodominant EBV peptide presented by HLA-B*08:01 demonstrated that the CDR1s and CDR2s made minimal energetic contributions to the formation of the TCR-pMHC complex, with both the CDR3s combining to dominate the “energetic landscape” (Borg et al., 2005).

When considering the individual residues that are involved in making contacts, it is worth taking each component of the TCR-pMHC complex in turn. With respect to the bound peptides, MHC binding motifs give potentially useful insights into the likely set of TCR-facing residues, with positions 2 and 9 (the anchor residues) least likely to be involved and the residues towards the middle of the peptide arguably the most likely, given that they are potentially in contact with CDRs even if the binding register of the TCR is somewhat unusual. With respect to the MHC-I molecule, the exposed residues of the MHC α-helices, particularly those on the “top” of the helices (facing towards the TCR), are by far the most likely to be in contact with the TCR. Owing to the diagonal crossing angle, the patch of TCR-binding residues is liable to be centred away from the mid-point of each α-helix, as shown in Figure 1.10. The suggestion that three specific MHC-I residues play an essential role in mediating MHC restriction (Tynan et al., 2005) has since been substantially discounted (Burrows et al., 2010).

1.4.3 HLA restriction

TCRs are said to be “MHC restricted” because they only bind to peptides when presented by MHC molecules. (Competing models that seek to explain MHC restriction are reviewed in (La Gruta et al., 2018).) But TCRs are also, to a significant extent, HLA restricted – that is, a single TCR will commonly bind to MHCs from a single HLA allele or a limited number of HLA alleles. Multiple studies have shown changes in TCR recognition when the same peptide binds to MHCs from closely-related HLA class I subtypes that differ by only one or two residues, for example, HLA-B*44:05 compared with HLA-B*44:02 and HLA-B*44:03 (Archbold et al., 2009) and HLA-B*57:01 compared with HLA-B*57:03 (X. G. Yu et al., 2007). Note that these studies do not contradict the observation that TCRs are cross-reactive, as published examples of cross-reactivity typically involve multiple peptides binding to MHCs encoded by the same HLA allele (see, for example, the Wooldridge study cited in section 1.4.1 (Wooldridge et al., 2012)).

However, the proportion of an individual’s TCRs that are HLA restricted is unclear, as is the extent to which they are routinely sensitive to minor differences between HLA subtypes. In a 2018 analysis of public TCRβ chains from a cohort of 666 healthy individuals, the authors identified 28 clusters of co-occurring TCRs, of which 26 were deemed HLA-associated (DeWitt et al., 2018). (Note that the authors often found it impossible to identify an association with a single HLA allele; they attribute this to the fact that certain HLA allele pairs “strongly co-occur across the cohort”, and in a subsequent evaluation of CMV-associated chains “did not detect a substantial degree of HLA promiscuity”.) The two clusters that were deemed “HLA-unrestricted” had distinctive properties, one appearing to consist of mucosal-associated invariant T (MAIT) cells that do not interact with pMHC (DeWitt et al., 2018). These results are interesting, but it is worth remembering that most TCRs are not public, and public TCRs may have sequences that are qualitatively different from those of other TCRs (see section 1.3.1).

1.5 Discussion

To conclude this chapter, it is worth considering some of the key points we have learned that have a direct bearing on the research presented later in this thesis concerning the prediction of CD8+ TCR-pMHC binding and CD8+ TCR-HLA associations.

Firstly, all the key molecular elements – MHC molecules, peptides and TCRs – are associated with exceptionally high levels of sequence diversity. The HLA genes that encode MHC molecules are among the most polymorphic in the human genome, with sequence diversity highest in the residues that form the MHC groove and the adjacent surfaces of its α-helices, i.e. those parts of the MHC molecules that are in contact with peptides and TCRs (see section 1.2.1). TCR chains are formed by a process called V(D)J recombination that is designed to promote sequence diversity (section 1.3.1), with sequence diversity highest in the CDRs – especially the CDR3s – that form contacts with the pMHC complex (section 1.3.2). Though short and somewhat constrained in sequence by the cleavage and trimming preferences of the MHC class I antigen presentation pathway (section 1.1.2), the antigenic peptides presented by MHC-I molecules are highly variable. In each case (MHC, TCR, peptide), there are many new sequences yet to be discovered.

Secondly, there are some important constraints. Certain key components have sequence length constraints: MHC $\alpha 1/\alpha 2$ domains (section 1.2.1), TCR CDRs (section 1.3.2), and antigenic peptides (section 1.2.3). There are also structural constraints: pMHC complexes are comparatively inflexible (although a degree of flexibility on TCR binding is sometimes observed – see section 1.4.1), bound peptides of the most frequently observed lengths are held in extended conformation (section 1.2.3), and most CDRs adopt a number of distinct conformation that have been assigned to canonical classes (see section 1.3.2), although most classes have been defined for unbound TCRs.

Finally, it is worth considering the implications of an important question that we are currently unable to answer: What proportion of TCR-pMHC complexes exhibit an orientation close to that illustrated in Figure 1.10. Even if one ignores “reverse polarity” binding, TCR crossing angles in solved structures are not confined to a narrow range (see section 1.4.2), but the number of solved TCR-pMHC structures

remains small, and there may be a bias towards unusual complexes. If it is the case that most TCR-pMHC complexes adopt similar binding geometries, there would likely be considerable overlap between the sets of residues involved in forming contacts between TCR and pMHC, even if there are no conserved binding motifs. The degree to which the prediction tasks addressed in this thesis are tractable is likely to depend on the answer to this question.

2 Computational Methods

2.1 Introduction

Given a TCR and a peptide-MHC-I complex, what one would like to predict is the likelihood that the TCR would be activated by that pMHC and generate an immune response. However, in practice, this requires information that is rarely (if ever) available concerning matters such as CD28 co-stimulation and signals from proinflammatory cytokines, so a more realistic aim is to predict whether the TCR is likely to bind to that pMHC.

One can divide approach for tackling this task into two broad categories — structure-based and sequence-based — although hybrid approaches that combine elements of both are feasible, one recent example being DeepAIR, which incorporates AlphaFold2 models (see below) with TCR sequence and gene usage information in an integrated deep learning framework (Zhao et al., 2023). Both structure-based and sequence-based approaches have potential pros and cons, although the balance between them is constantly changing as new prediction methods are developed and new data becomes available.

In principle, a structural approach to peptide-MHC-I binding prediction has the potential to model the 3D conformations and flexibility of the molecules involved, notably the TCR's CDRs. Given separate solved structures of TCRs and pMHC complexes, docking methods – both generic protein docking tools (Peacock & Chain, 2021) and TCR-specific docking tools such as TCRFlexDoc (Pierce & Weng, 2013) – have achieved some success in identifying whether and how a given TCR and pMHC fit together. However, there are currently only 634 PDB structures containing TCRs in the curated STCRDab database (Leem et al., 2018) (https://opig.stats.ox.ac.uk/webapps/stcrdab-stcrpred, 09/09/2024), of which 173 are human αβTCRs bound to MHC class I molecules with a resolution better than 3.5Å. If a structure-based approach is to become broadly applicable, what is needed is a way of accurately modelling the structures of TCR and pMHC, and preferably the entire TCR-pMHC complex.

The 2020 performance of AlphaFold 2 (Jumper et al., 2021) at CASP14 has been widely recognised as a transformative moment in the field of protein modelling. In a subsequent version called AlphaFold-Multimer (Evans et al., 2021), it was extended to handle multi-chain protein complexes. Various attempts have been made to model TCRs and/or pMHCs using: these tools themselves, AlphaFold 2 (D. Wu et al., 2024) and AlphaFold-Multimer (e.g. (Abanades et al., 2023)); adaptions of both tools, such as TCRmodel2 (Yin et al., 2023) and TCRdock (Bradley, 2023); and novel deep-learning methods specifically designed to model TCRs or MHCs, such as MHCFold (Aronson et al., 2022). A recent survey of such initiatives by McMaster and co-workers concluded that all these approaches have their limitations, with several contributory factors identified, ranging from the lack of appropriate experimental data for tool development to the inherent challenges of modelling the weak binding affinities and changes in CDR loop conformation that characterise TCR-pMHC binding (McMaster, Thorpe, Ogg, et al., 2024) .

The key advantage of sequence-based approaches to the TCR-pMHC binding prediction task is the much higher volumes of available data – at least tens of thousands of TCR-pMHC complexes with paired TCR α and β chains (the challenges involved in acquiring a clean set of such data with complete TCR sequences is discussed in section 3.2.2), the MHC sequences for many thousands of HLA alleles and billions of TCR sequences with unknown targets, both paired and unpaired. The fundamental challenge for sequence-based approaches is that the formation of TCR-pMHC complexes entails the engagement of three-dimensional elements involving different combinations of residues depending on the complex. In order to make accurate predictions, sequence-based approaches are likely to be more tightly constrained to tackling complexes that are similar to those in the training set compared to structure-based approaches. But the degree of similarity required, particularly with modern deep learning approaches, remains an open question, as does the extent to which current sequence data is (at least to a useful degree) representative of the denser parts of TCR-pMHC space – although it certainly has to be acknowledged that, to quote a recent review, we currently have access to " just a minute fraction of the possible sample space of TCR–antigen binding pairs" (Hudson et al., 2023).

The research in this thesis focuses on sequence-based methods and is in two parts: Chapter 3 covers the benchmarking of several deep learning TCR-pMHC binding

prediction methods, and Chapter 4 presents a novel deep learning method for predicting TCR-HLA associations. The remainder of this chapter provides: a short introduction to other sequence-based methods for TCR-pMHC binding prediction (distance metrics, clustering and conventional machine learning in section 2.2); a more detailed introduction to deep learning (section 2.3) and transfer learning approaches (section 2.4); a discussion of data-related challenges (section 2.5); and a brief guide to the computational resources and platforms used in this research (section 2.6).

2.2 Distance metrics, clustering and machine learning

This section sets the context for the introduction to deep learning that follows in section 2.3. None of the methods mentioned here features in the benchmarking performed in Chapter 3. However, TCRdist3 (section 2.2.1) was used in the correlation analyses presented in sections 3.5.2 and 4.5.

2.2.1 Distance metrics and clustering

In clustering approaches to TCR-pMHC binding prediction, TCRs with similar characteristics (inferred from sequence similarity) are grouped together based on the underlying hypothesis that similar TCRs will bind to similar pMHCs. Several metrics have been devised to measure the distance between TCRs. One of the more sophisticated methods is TCRdist (Dash et al., 2017), which has since been adapted for use in TCRMatch (Chronister et al., 2021) and revised in TCRdist3 (Mayer-Blackwell et al., 2021). The latter compares the sequences of all the CDRs (including the CDR2.5s) belonging to a given pair of TCRs, having pre-trimmed the highly conserved N- and C-terminal residues from the CDR3s. Distances are calculated by comparing the corresponding CDRs of the two TCRs: substitution penalties are awarded using BLOSUM62; insertions are awarded the same high penalty as non-conservative substitutions; the number of CDR3 penalties is multiplied by a factor of three; and the final distance is the sum of all penalties (Mayer-Blackwell et al., 2021).

Various research groups have used TCRdist3 and other distance metrics to build clusters of similar TCRs. The authors of TCRdist3 used the metric to identify public clusters of TCRs directed against SARS-CoV-2 (Mayer-Blackwell et al., 2021). In

another study, TCRs were clustered using a tool called GLIPH2 to identify *Mycobacterium tuberculosis*-specific CD4+ TCRs (Huang et al., 2020).

Although complex distance metrics such as TCRdist3 that have been optimised for measuring TCR-specific similarities are potentially useful (indeed, TCRdist3 is used in the correlation analyses presented later in the thesis), it is reasonable to assume that the relationship between measurements of the distance between TCRs and differences in the function of these TCRs is complex and nonlinear. This provides a key motivation for using machine learning algorithms to address these challenges.

2.2.2 Machine learning

Machine learning methods have become an increasingly popular way of addressing complex biological tasks, including TCR-pMHC binding prediction. (Deep learning is a type of machine learning but is considered separately in section 2.3.)

TCR-pMHC binding prediction is commonly treated as a binary classification task ("Does a given TCR bind to a given pMHC, true or false?"). Methods are trained using labelled data, where each input pattern (a description of a TCR-pMHC complex) is paired with the desired output ("binds" or "does not bind"). After training, the methods can be used to make predictions for unlabelled data (i.e. the input is known but the desired output is unknown).

Two specific machine learning algorithms have been used to address this task: random forest (RF) and gradient-boosted trees (GBT). RF (Ho, 1995) tackles a classification task by building an ensemble of decision trees. A decision tree is essentially a set of conditions arranged in a binary tree (i.e. a tree in which each non-leaf node has exactly two children), where a condition might be "Is the current HLA allele HLA-A*02:01?", or "Does the current peptide have a Tyrosine at position 9?". After applying many decisions, a prediction is made. A single decision tree will tend to overfit to the training data; RF (partially) overcomes this problem by generating multiple decision trees from different subsets of the training data and then classifying unseen data based on the combined output from this ensemble of trees. The TCR-pMHC binding prediction tools TCRex (Gielis et al., 2019) and epiTCR (Pham et al., 2023) use the RF algorithm implemented in the Python library scikit-learn (Pedregosa et al., 2011).

Whereas each RF decision tree tends to overfit to the training data, GBT methods such as XGBoost (Chen & Guestrin, 2016) build shallow trees, each of which underfits to the training data. Trees are added sequentially, with each new tree trained to address the errors made by the current collection of trees. With GBT, the risk of overfitting to the data is associated with the addition of too many trees, rather than the individual trees themselves. The TCR-pMHC binding prediction tool SETE (Tong et al., 2020) uses the GBT algorithm implemented in scikit-learn and treats TCR-pMHC binding prediction as a multi-class problem, rather than binary classification (see below).

A key issue for any predictive model applied to the TCR-pMHC binding prediction task is to decide how to represent input information about a TCR-pMHC complex, and it is here that traditional machine-learning methods have limitations in terms of their ability to handle long protein sequences of variable length. Both epiTCR and SETE represent a TCR using its CDR3β sequence alone, whereas TCRex adds the CDR3α sequence and provides categorical labels for the TCR V and J genes. Only epiTCR encodes the epitope sequence; instead, TCRex comes in multiple versions, each of them epitope-specific, and SETE treats a limited number of epitopes as targets (hence multi-class prediction) and encodes them categorically. Only epiTCR incorporates HLA allele information, which is achieved using the fixed-length pseudo-sequence approach described in section 3.2.2. Note that categorical encoding is appealing here because it can be used to represent long and/or variable length data in a concise fixed-length format. However, there is a fundamental drawback – all categorical labels are, in effect, equally distant from each other, so all information about the similarities between particular categories is lost.

2.3 Deep Learning for TCR function prediction

Deep learning is a subset of machine learning methods that is based on artificial neural networks (ANNs). The building blocks of an ANN are nodes or artificial neurons, so named because they are loosely inspired by the neurons of the human brain. Nodes are connected by weights. In the simplest form of an ANN, a feed-forward neural network (FNN), the output of a node is the result of a nonlinear function (called the activation function) being applied to the weighted sum of the

node's inputs. Training an FNN is an iterative process whereby a set of input patterns with desired outputs (the training set) is presented to the network, the errors in the network outputs are calculated, and the weights are adjusted to reduce the total error for all patterns, commonly via a process known as backpropagation.

FNNs are one of the most popular architectures in conventional machine learning, but also provide the basic architecture for an important type of deep neural network, the convolutional neural network (CNN).

2.3.1 Convolutional neural networks

The convolutional neural network (CNN) – can be regarded as the breakthrough architecture in deep learning (for the advances made in the field of image processing) and has been used in several TCR-pMHC prediction tools, either on its own or as a part of a more complex model. The CNN architecture comprises three distinct types of layers: a convolutional layer, a pooling layer and a fully connected layer. A convolutional layer generates feature maps, each capturing information about a particular repetitive feature irrespective of its location (within an image or sequence). A pooling layer subsamples the feature maps, thereby reducing the risk of overfitting to the training data. A typical CNN may have multiple convolutional layers interspersed with pooling layers, after which come one or more fully connected layers. The job of the fully connected layers is to make predictions based on the features they receive as inputs in much the same way as a conventional FNN.

The most famous applications of CNNs are in image recognition, where the inputs comprise a 2-dimensional grid of pixels, each consisting of three values (red, green and blue). This has been straightforwardly adapted to handle protein sequences, which can be viewed as a 1-dimensional "image" with 20 values (the standard amino acids) at each position (equivalent to a pixel). Four of the tools benchmarked in Chapter 3 use CNNs: pMTnet (Lu et al., 2021), NetTCR-2.0 and 2.1 (Montemurro et al., 2021, 2022) and TEINet (Jiang et al., 2023).

However, CNNs have limitations when handing long protein sequences, as the network needs to handle a complete sequence as a single input – big enough to accommodate the longest sequence in the dataset; lots of inputs means lots of network weights and, in general, the more weights a network has, the more data is

needed to reduce the risk of overfitting and the more internal calculations need to be performed, increasing training times. In such cases, there is a tendency to adopt a more aggressive pooling strategy, which can lead to the loss of useful information.

2.3.2 Recurrent neural networks

An alternative to treating a protein sequence in a similar way to an image is to treat it as a series of amino acid residues that one inspects one at a time in a similar way to time-series data (e.g. changes in temperature, heart rate or stock prices) or natural language (e.g. the words within a sentence). Such an approach has two practical advantages: each input is much smaller (i.e. a single residue) and sequences of different lengths have no impact on the dimensions of the network.

The basic architecture needed to handle data in this way is called a recurrent neural network (RNN), which is essentially the same as an FNN but with an added hidden state mechanism – often described as the RNN's "memory" capability – for storing information about previous elements (e.g. amino acids) in the current sequence.

However, a key limitation with a standard RNN is that, as the sequence gets longer, the amount of information retained about elements that appeared early in the sequences gets smaller. Hence, RNNs are said to have only "short-term memory", whereas it is often important to capture long-range dependencies (e.g. between words far apart in a sentence or residues far apart in a protein sequence). The long short-term memory (LSTM) (Hochreiter & Schmidhuber, 1997) is designed to address this problem by allowing the network to learn what it needs to remember (irrespective of where it occurs within the sequence) and what it can afford to forget. ERGO II (Springer et al., 2021) and pMTnet (Lu et al., 2021), two of the TCR-pMHC binding prediction tools benchmarked in Chapter 3, incorporate an LSTM. A third method, TCellMatch (Fischer et al., 2020), uses a bidirectional gated recurrent unit (GRU) model; a bidirectional RNN (Schuster et al., 1997) is one that proceeds through the sequence of elements in both directions, and a GRU (Cho et al., 2014) can be viewed as a simplified version of an LSTM.

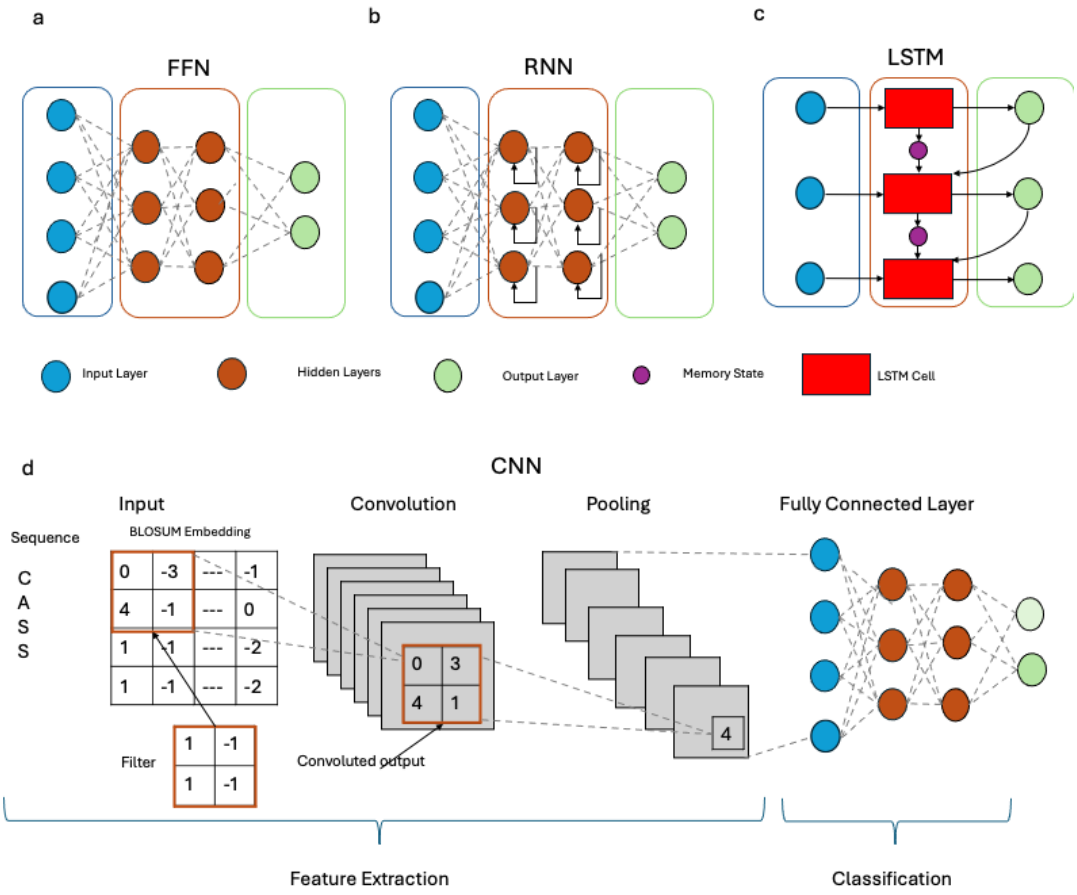

Figure 2.1 Deep Learning Architectures. A neural network contains 3 main layers – an input layer, a hidden layer and an output layer. **a.** A simple feed forward network processing information in one direction. **b.** Recurrent Neural Network: An RNN stored the previous output in the memory and uses it as in additional input to the next state. **c.** An LSTM maintains a memory state to store important information for all time steps. **d.** CNN architecture: In this the input sequence is first converted into a numerical embedding using BLOSUM62 matrix resulting into a 5x20 matrix. Next the convolution layers use the filters to extract the features. It is then passed through pooling layers. A fully connected layer then integrates the features and map to the output layer.

2.4 Transfer Learning

2.4.1 Introduction

As noted in section 2.1, the amount of TCR-pMHC sequence data (tens of thousands of samples) is dwarfed by the amount of unlabelled TCR sequence data (billions of samples), i.e. data from TCRs where there is no information about their potential antigenic targets. Unlabelled TCR sequence data contains a lot of potentially useful information about the relationships within and between TCR sequences.

A strategy that enables one to exploit large quantities of unlabelled data is called transfer learning, where a preliminary model is trained on bulk unlabelled data (pre-training) and then repurposed for the main task of interest (fine-tuning). The central concept behind transfer learning is that the pre-trained model should learn the "underlying grammar" of the domain from the input data. The term "underlying grammar" betrays the preeminent application area for transfer learning, namely within the field of natural language processing, but the term has also been adopted in the biological domain, for example, "we use unlabeled TCR sequences to learn the underlying grammar of the naturally occurring TCR sequence space" (K. Wu et al., 2021).

Several transfer learning models have been developed that use bulk collections of protein sequences, notably ProtTrans (Elnaggar et al., 2021), trained on the Big Fantastic Database (BFD) (Steinegger et al., 2019). Two models are of particular relevance to this thesis: the TCR-BERT model (K. Wu et al., 2021), developed using a dataset of 88,403 paired TCR sequences, and the STAPLER model (Kwee et al., 2023) using nearly 160,000 CDR3 sequences and over 180,000 antigenic 9-mers.

2.4.2 Transformers

With long sequences (e.g. of words or amino acids), LSTMs struggle to capture long-range dependencies because they process them sequentially, one element at a time, and encoded information about the whole sequence has to be compressed into a fixed-length vector. A transformer (Vaswani et al., 2017) is a deep learning architecture that addresses the preceding issues by a) scanning the whole sequence at each step and b) by assigning attention weights to tokens (e.g. words or residues)

based on their contribution to the transformer's output (tokenisation is addressed in the next section). Several different types of attention mechanisms are possible: self-attention, which focuses on relationships between tokens within a single sequence; cross-attention, which focuses on relationships between different sequences (e.g. between sequences from a TCR and an MHC molecule); and multi-head attention, which allows multiple dependencies to be captured for a single token.

Several existing transformer-based models are relevant to this thesis. BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), a notable state-of-the-art natural language processing (NLP) Transformer model, has inspired two BERT-based models that address the TCR-pMHC binding prediction problem: TCR-BERT (K. Wu et al., 2021) and TCRBert (Yoo et al., 2024). The TPBTE transformer model (J. Wu et al., 2023) incorporates cross-attention between a TCR and an epitope in making TCR-pMHC binding predictions. Lastly, ATM-TCR (Cai et al., 2022) uses a multi-head attention to predict TCR-pMHC binding affinity.

In this research, I have used a variant of BERT called DeBERTa (He et al., 2020) (4.3). DeBERTa uses a so-called "disentangled attention mechanism" whereby each token is represented by two vectors, one encoding its position and the other its contents. In the present context, this enables the model to work out whether, for example, it is the position of a given residue or its amino acid type that is more important.

2.4.3 Tokenisation

Tokenisation is the process of converting input data into tokens that are ultimately converted into the numerical representation required for the computation to take place. Much of this can be automated (e.g. using DeBERTa Tokenizer), but there are also choices to be made about what features will be treated as tokens. In the present context, the minimal and standard choice is to tokenise the individual amino acid residues. However, in this research, additional tokens were created for higher level features, notably a separate token for each of the different types of CDR (including the CDR2.5s) and for the sequence of the MHC molecule.

More details are given in Appendix 1.

2.4.4 Pre-training

The aim of pre-training is to train a model using a large set of unlabelled data (in the present context, TCR and MHC sequences) to address a preliminary task; the task should be one that will enable the model to learn the properties of the data that will subsequently prove beneficial when the model is transferred to the main task of interest (in the present context, the prediction of TCR-HLA associations). In this research the preliminary task involved using the masked language modelling (MLM) technique, which is widely used for training transformer models. Within a given TCR or MHC sequence, 40% of the residues were masked at random, and the model learned to predict the amino acid types of the masked residues. Such a task encourages the model to learn about the relationships between residues at different positions and between residues of different amino acid types.

As there are no known associations between TCR and MHC sequences in the unlabelled dataset, the model was trained using a single TCR or MHC sequence at a time (both TCRs and MHCs are identified by their own domain-specific custom tokens). In pre-training, the DeBERTa model uses the Adam optimiser to minimise the overall prediction error by adjusting the model parameters.

2.4.5 Fine-tuning

Having pre-trained a model, it can be adapted to address a specific prediction task of interest. This involves adding one or more additional layers. For the TCR-HLA association prediction task, I used the ‘‘DeBERTaForSequenceClassification” class, to add a classification layer. At this stage the model was trained as a cross-encoder, with the sequences for the CDRs and MHC concatenated together as a single input. Only the new layer underwent modification.

The DeBERTa TCR-HLA association predictor was fine-tuned multiple times on different datasets, as described in Chapter 4.

2.5 Data-related issues

Central to the research presented in this thesis is the creation of a TCR-pMHC dataset from public resources, as described in section 3.2, and subsequently repurposed for the TCR-HLA associations task, as described in section 4.2. This section briefly addresses some underlying issues with TCR-pMHC binding data that will have undoubtedly impacted this research.

Most of the collected data was derived from experiments using MHC multimers. Over 80% of the records stored in VDJdb (Shugay et al., 2018) involved MHC dextramers, but the type of MHC multimer is not specified for the records stored in all resources, and an unknown proportion of the records in this research will have come from experiments using MHC tetramers (the second most common source of records in VDJdb). This is problematic because the binding affinity threshold for class I MHC tetramers is "significantly higher" (to quote Dolton et al.) than the CD8+ T cell activation threshold, implying a bias towards higher-affinity TCRs. This problem does not arise with dextramers (Dolton et al., 2014).

A recent review highlights several potential issues with 10x Genomics data, including a bias towards high-specificity TCRs and concerns over non-specific binding (Hudson et al., 2023). Some 10x data was incorporated in the datasets used here, although efforts were made to remove the most unreliable samples from this source (see section 3.2.2).

Some resources, such as IEDB (Vita et al., 2019), contain data accumulated over decades using a variety of different assays. The relative sensitivity and reliability of these different methods is hard to judge, as is the potential bias they may introduce, but there is at least one systematic comparison that highlights some of the potential issues (Y. Sun et al., 2003).

Apart from issues related to TCR binding affinity, TCR-pMHC binding data is associated with several additional biases, notably those concerning the lack of data diversity with respect to HLA alleles, epitopes and the sources from which those epitopes derive. A recent analysis of various datasets, including some used in this research, concluded that they are skewed towards "the 3-6 most common HLA alleles" and "a limited number of epitopes, predominantly of viral origin" (Weber et

al., 2024). This is indeed reflected in the datasets used in this research, which have a strong bias towards a small number of CMV epitopes and two HLA alleles, HLA-A*02:01 and HLA-A*03:01 (see Figures 3.2 and 4.1).

2.5.1 Negative Data

There is a limited quantity of experimentally validated negative TCR-pMHC binding data (i.e. cases where a given TCR is known not to bind to a given pMHC) in public resources. Given that some T cell assays have binding affinity thresholds that are not sensitive enough to pick up weak CD8+ TCR binders (as noted in the opening section of this chapter), there is also a risk that some samples labelled as negative are, in fact, undetected positives.

In the absence of sufficient experimentally validated negative data, it is common practice to generate negative pairs by randomly mismatching TCRs with pMHCs. Given the propensity of TCRs to cross-react (see section 1.4.1), there is clearly a risk that any such mismatched pair might in fact be a positive, although the rate at which these false negatives occur is assumed to be low.

Two different strategies have been implemented for generating negative data by mismatching TCRs with pMHCs. The first involves mismatching TCRs with pMHCs drawn exclusively from the positive dataset used for training and testing. Depending on the desired ratio of negatives to positives and the epitope imbalance within the dataset, creating sufficient negative samples may be problematic. The second approach is to randomly select TCRs from a different source. For example, Gao and co-workers selected (for a CDR3β-only prediction method) TCRs from the repertoires of 587 healthy patients (K. Wu et al., 2021) (Gao et al., 2023). However, it has been demonstrated multiple times that this type of approach may artificially boost the performance scores of a predictive tool because it is able to learn features of the TCR sequences that differ between the positive and negative sets and thereby make correct predictions without reference to the peptide (Grazioli et al., 2022; Moris et al., 2021).

The first of these strategies was the one adopted for the research presented in this thesis (see sections 3.2.3 and 4.2.1).

2.6 Computational Resources

2.6.1 Open-source platforms

The pre-trained models used in this research for TCR function prediction are stored on an open-source data science and machine learning platform called Hugging Face[6] Hub. The platform's Trainer class was used to train (during both pre-training and fine-tuning) the model that was developed for the TCR-HLA associations prediction task (see Chapter 4). The Trainer class uses the OPTUNA[7] framework for hyperparameter optimisation, notably the learning rate and number of training epochs.

One key advantage of Hugging Face is that it has an open-source framework called Transformers that provides support for transformer model development. It was used extensively for the research presented in this thesis.

2.6.2 Cloud computing

Highly complex Natural Language models require significant computational resources for building, training and testing. Cloud computing offered on-demand access to a shared pool of configurable computing resources and affording a way of avoiding the queuing time associated with HPC clusters. During my PhD, I made a successful application for an Oracle for Research grant, which allowed me to parallelise many of the training tasks and significantly expanded the scope and depth of model training and testing.

[6] https://huggingface.co/
[7] https://optuna.org/

Table 2.1 Computational resources that were used in this research.

Name	GPUs Specification	TPU Specification	Cost
AWS (Amazon Web Services)	8 x A100 (one of the available GPU instances)	-	£28/hour
Birkbeck College servers	8 x NVidia GeForce	-	-
Google Research TPU Cloud	-	5 x on-demand Cloud TPU v3 devices; 5 x on-demand Cloud TPU v2 devices; 100 preemptible Cloud TPU v2 devices	30 days free trial
Oracle For Research	2 x NVIDIA Tesla A10	-	£45,000 Grant for 12 months

2.7 Discussion

This chapter has set the context for the computational research presented in chapter 3 and 4. It began by engaging with the broader context: what computational approaches can be used to address the TCR-pMHC prediction task and what are some of the key challenges? One of those challenges concerns the availability of good quality data. This issue is taken up subsequently in section 2.5 and will be revisited again in section 3.2.

The other major focus of the chapter was on computational methods – both the various strategies used by others to address the TCR-pMHC prediction task (ranging from clustering to advanced deep learning strategies), and the DeBERTa-based transformer I developed for predicting TCR-HLA associations. Here the main motivation has been to explain some key differences between the various approaches and to provide a broad rationale for the choices made during this research (without going into technical details).

This chapter ends with a brief insight into how access (or, at times, the lack of access) to Cloud-based computation resources played a vital role in this research which mainly focuses on developing a predictive model to answer one key question for TCR-antigen specificity, i.e. Can the given TCR of unknown antigen specificity bind to the given pMHC complex?

3 Benchmarking TCR-pMHC Binding Prediction Tools

3.1 Introduction

In the past few years, many TCR-pMHC binding prediction tools have been released, varying in the architectures and algorithms used, in the selection of biological features and data encoding strategies. Tool performance is reported (sometimes using different evaluation metrics) based on datasets that may vary significantly. Importantly, some datasets contain negative data that is associated with misleadingly high levels of performance (as discussed in section 2.5.4), and evaluations typically involve at least a subset of TCRs that are known to bind to pMHC complexes that are present in the training set. Hence, it is not feasible to make a useful comparison of the ability of these methods to generalise (i.e. assess their ability to predict whether a given TCR will bind to peptides that are not present in the training data), from the performance levels reported by the tools' authors.

In this thesis, the emphasis is firmly on the ability of tools to generalise. Given that only a small fraction of antigenic peptides are present in public datasets, the ability for a tool to generalise is a key measure of its potential usefulness, as has been widely acknowledged (Castorina et al., 2023; Deng et al., 2023; Montemurro et al., 2022). Considering the vast combinatorial space of potential TCR-pMHC combinations, and the various processes that lead in practice to the creation of novel antigenic peptides – including the emergence of novel zoonotic diseases (e.g. SARS-CoV-2, Ebola, avian influenza), the rapid evolution of certain established viral pathogens (e.g. influenza A, HIV, norovirus), and the formation of neoantigens by tumour cells – it seems reasonable to assume that the ability to make generalised predictions of TCR-pMHC binding will remain an important goal for the foreseeable future. Equally, given the complexity of TCR-pMHC interactions and the poor coverage of TCR-pMHC space in public datasets suggests, it appears reasonable to assume that the ability of tools to generalise is likely to be constrained to regions within that space where there is scope to learn what might be termed "transferable binding patterns", although it is unclear how such patterns may be characterised.

This chapter focuses on: the sources of data available for tool evaluation and the creation of a benchmark dataset (section 3.2); available TCR-pMHC binding prediction tools, and the selection of a subset of these tools for benchmarking (section 3.3); the choice of cross-validation strategy and performance metrics (section 3.4); and the results of the benchmarking exercise together with a correlation analysis exploring the relationship between properties of the data and predictive performance (section 3.5). But first, it is worth considering what concurrent benchmarking efforts have been undertaken in other research.

3.1.1 Other benchmarking initiatives

Several previous comparative assessments of TCR-pMHC binding prediction tools have been undertaken, each with important differences to the assessments presented here. Arguably the most important of these earlier assessments was the ImmRep 2022 community benchmarking initiative, with tool developers invited to participate in a workshop, and datasets made available for training and testing. The workshop's aim was to "evaluate and compare the obtained outputs to classify the approaches and most importantly, to help identify an ideal dataset and optimal evaluation strategies for future follow-up efforts" (Meysman et al., 2023). A total of 23 prediction models were evaluated during the workshop.

The main difference between ImmRep 2022 and the benchmarking undertaken here is that the former focused exclusively on making predictions for TCR-pMHC complexes where the peptide is present in the training set (bound to at least one and often many distinct TCRs, none of which are present in the test set) – described as "the 'seen' epitope setting" (Meysman et al., 2023) – whereas the benchmark here focuses on generalisation to unseen peptides. Key lessons of the ImmRep 2022 initiative are described as follows in the Abstract of the key publication arising from the workshop: "the use of paired-chain alpha-beta, as well as CDR1/2 or V/J information, when available, improves classification obtained with CDR3 data, independent of the underlying approach" (Meysman et al., 2023). Although it might appear reasonable to assume that these lessons also apply to the performance of tools in the "unseen" epitope setting, this should not be taken for granted and additional factors are likely to be important, as investigated in this thesis.

To my knowledge, only two benchmark evaluations focusing on the generalised (unseen peptide) TCR-pMHC binding prediction task have been published so far (Grazioli et al., 2022), (Deng et al., 2023) with the former confined to only two predictors and the latter confined to the most common HLA allele, HLA-A*02. Further information about these existing benchmark studies will be covered later in this chapter, but it is worth noting that there is a widespread belief within the field that, to quote the authors of one of these recent benchmark evaluations, “modern deep learning methods fail to generalize to unseen peptides” (Grazioli et al., 2022); the authors of the second such benchmarking study draw the same conclusions (Deng et al., 2023).

3.2 Benchmark Datasets

This section focuses on the creation of benchmark datasets suitable for the evaluation of TCR-pMHC binding prediction tools, with an emphasis on their ability to generalise to unseen peptides.

Before considering the details, it is worth making two broad points. The first arises from the fact that many TCR-pMHC samples available in public databases only provide information about the TCR β chain and commonly just the CDR3β. Consequently, there is a trade-off between the number of TCR-pMHC samples and the amount of information – notably about the TCR – available about each sample. On the one hand, given the observations of Meysman and co-workers (quoted in section 3.1.1) concerning the predictive improvements observed in the “seen” epitope setting with the addition of information about both TCR chains and all their CDRs (Meysman et al., 2023), it appears highly desirable to incorporate such information in a benchmark dataset designed to evaluate performance on the harder, unseen peptide task. On the other hand, the performance of machine learning algorithms often improves with the amount of training data (up to some problem-/tool-specific threshold). The decision taken here was to incorporate the maximal amount of information about TCRs within the dataset, while also expending a good deal of effort to maximise the number of samples by integrating data from multiple sources and imputing missing information where possible (as discussed below in sections 3.2.1 and 3.2.2). Given the extent of the work involved, the effort was split

with a second member of Prof Adrian Shepherd's research group, Justin Barton; decisions about what data to download and how to process it were taken collectively.

Secondly, given that different prediction tools utilise data features in different ways (e.g. some represent the MHC molecule within a sample using a categorical encoding of its HLA allele name, whereas other encode parts of its amino-acid sequence), there is a need for multiple versions of a given benchmark dataset. Some of the implications of this are addressed in section 3.2.2, although most of the details are deferred until the discussion of benchmark tool selection in section 3.3.

3.2.1 Sources of binding data

The benchmark dataset was created using TCR-pMHC binding data selected from five public databases and repositories: VDJdb (Shugay et al., 2018), McPAS-TCR (Tickotsky et al., 2017), TBAdb (Wei Zhang et al., 2020), IEDB (Vita et al., 2019), and 10x Genomics public datasets (10X Genomics, Pleasanton, CA, USA; www.10xgenomics.com/datasets/). Although there is a degree of overlap between these databases, each contributed a sufficient number of unique samples to the final benchmark dataset to make the inclusion of all four worthwhile. Each resource will now be considered in turn.

VDJdb (Shugay et al., 2018) is a curated database of T cell receptor sequences with known antigen specificities. It was downloaded on 06/01/2023 from https://vdjdb.cdr3.net/ with selection criteria: human TCRs with paired α and β chains associated with MHC I molecules.

TBAdb (Wei Zhang et al., 2020) is a manually curated database containing both TCRs and B cell receptors with known antigenic targets that formed a discrete deposition within the Pan Immune Repertoire Database (PIRD). It was downloaded in its entirety on 05/05/2020 from https://db.cngb.org/pird/. (At the time of writing, TBAdb is not available via the preceding link but can be downloaded from https://gitlab.com/immunomind/immunarch/-/tree/master/private?ref_type=heads.)

The Immune Epitope Database (IEDB) (Vita et al., 2019) is a large public database that stores information about B cell and T cell epitopes derived from experimental data. Data was downloaded from IEDB on 15/05/2023 from https://www.iedb.org/ with selection criteria: T cell assay data associated with MHC class I linear epitopes,

human host and any disease. Both “calculated” data (in which CDR3 locations are inferred via an automated analysis of TCR sequences) and “curated” data (in which CDR3 locations are assigned manually) were merged to maximise the number of complete records.

McPAS-TCR (Tickotsky et al., 2017) is a manually curated database containing human and mouse TCR sequences with their respective antigens (associated with various pathologies). The most recent update of the database, dated 10th September 2022, was downloaded in its entirety from https://friedmanlab.weizmann.ac.il/McPAS-TCR/.

From 10x Genomics, four BEAM-T and four multiplex datasets were downloaded on 19/03/20203 from https://www.10xgenomics.com/datasets, these being the only human CD8+ T cell datasets available at that time. All were derived from PBMC samples. The BEAM-T datasets were associated with EBV, CMV, influenza and SARS-Cov2. The four multiplex datasets were from healthy donors.

3.2.2 Data processing

Prior to combining data from the different sources, the downloaded data was filtered and cleaned.

The TBAdb and McPAS-TCR datasets were filtered to include only human TCR with paired chains. Some additional TBAdb entries were removed that had missing peptide or HLA information.

With respect to the data downloaded from VDJdb, mislabelled samples were removed, notably entries that (in spite of the selection criteria applied when the data was downloaded) contained MHC allele names that are not associated with human class I (notably the mouse allele “H-2Kb” and names containing the class II-specific string “DRB”). α and β chain V gene annotations that contained only the gene name (e.g. TRBV10-3) were manually assigned the appropriate allele name (e.g. TRBV10-3*02) by imputation, for example, by inferences derived from a comparison of the available sequence information (notably the sequence forming the start of the CDR3) with the canonical TCR reference sequences available from IMGT (Manso et al., 2022)(URL: https://www.imgt.org/vquest/refseqh.html).

With respect to the data downloaded from IEDB, both “calculated” data (in which CDR3 locations are inferred via an automated analysis of TCR sequences) and “curated” data were merged to maximise the number of complete records, with missing values in the “calculated” set (e.g. the V and J alleles of TCR α and β chains) imputed from the “curated” data where possible. For IEDB entries with missing MHC information, data from VDJdb, McPAS-TCR, and TBAdb were used to impute the relevant MHC allele where possible. In all cases where the imputation of essential information proved impossible, the relevant entries were removed from the dataset.

A subset of entries from several resources had MHCs specified only to the allele group level (e.g. HLA-A*02). Such MHCs were manually assigned to a specific HLA protein level (e.g. HLA-A*02:12) based on MHC-peptide binding data from other sources, notably the training set provided with NetMHCpan 4.1 (Reynisson et al., 2020). Where no such evidence was found, allelic assignments were made to the allele that occurs most commonly in global populations according to the Allele Frequency Net Database (Gonzalez-Galarza et al., 2020) (e.g. in the case of allele group HLA-A*02, the most frequent protein-level allele is HLA-A*02:01).

With respect to the data downloaded from 10x Genomics, cells were filtered to: include only productive, high-confidence, full-length $\alpha\beta$ T cells with completely specified V and J genes; and exclude those not containing exactly two distinct TCR chain sequences (thereby removing dual-receptor T cells and experimental artefacts such as doublets and GEMs without cells). As a final step, the gene expression data associated with a given 10x experiment was used as the basis for identifying a subset of TCR-pMHC complexes considered to be “binding”, as follows: a given TCR-pMHC complex was considered to be “binding” if the associated peptide UMI count was greater than three standard deviations higher than the UMI count for the experiment’s negative control peptide. (Note that this approach to filtering 10x Genomics data, devised by Justin Barton in consultation with other members of Prof Shepherd’s research group, is simpler than the subsequently published ITRAP method (Povlsen et al., 2023); when applied to the 10x Genomics data used in this research, over 98% of the TCR-pMHC complexes selected by the two approaches were the same.)

After converting the relevant fields of all datasets into a standardised format, the datasets were combined into a single set and duplicates were removed. Two final

filters were applied to this combined dataset: the first removed entries with peptides shorter than 8 or longer than 12 amino acids, and the second removed records with non-standard or ambiguous amino-acid types.

Several of the TCR binding data repositories store only a minimal description of a given TCR consisting of its CDR3 sequences and its V and J gene identifiers. However, certain TCR-pMHC prediction models require either full-length TCR amino acid sequences or complete sets of their CDR sequences. In such cases, the Python tool Stitchr (Heather et al., 2018) was used to reconstitute the complete α and β chain amino acid sequences from their minimal descriptions. To specify the CDR1 and CDR2 start and end residues within a reconstituted sequence, the antibody and TCR sequence numbering tool ANARCI (Dunbar & Deane, 2016) was used in combination with the CDR positional definitions specified by IMGT (Manso et al., 2022).

Several TCR-pMHC prediction models require HLA pseudo-sequences of the form devised by Nielsen and co-workers at DTU Health Tech to represent the MHC components of the TCR-pMHC complexes. Such pseudo-sequences consist of 34 amino acid residues that are putative contact residues (occurring "within 4.0 Å of the peptide in any of a representative set of HLA-A and -B structures with nonamer peptides") and known to be polymorphic (Nielsen et al., 2007).

Finally, an analysis of the remaining samples in the dataset showed that approximately 50% of the positive samples were associated with a single human cytomegalovirus (CMV) epitope. This was considered undesirable as there was a risk that predictors trained on such a dataset would specialise in making prediction for this single epitope. Consequently, it was decided to randomly downsample this epitope from around 14,000 samples to 4,200.

The final number of positive samples in the dataset was 16,220. The positive samples in the benchmark dataset breakdown in terms of epitopes is given in Figure 3.1, and for HLA alleles in Figure 3.2. For comparison with other benchmark studies, the TChard dataset (Grazioli et al., 2022) is much larger, but most of the samples are CDR3β only. The Deng dataset (Deng et al., 2023) contains 15,331 positives – nearly as many as our own – despite being confined to HLA-A*02. However, whereas the Deng dataset only contains CDR3s (paired α and β), ours contains full TCR α and β sequences; in cases where it proved impossible to acquire a full sequence

using Stitchr or the boundaries of CDRs 1 and 2 using ANARCI (as described above), the sample was removed from our dataset.

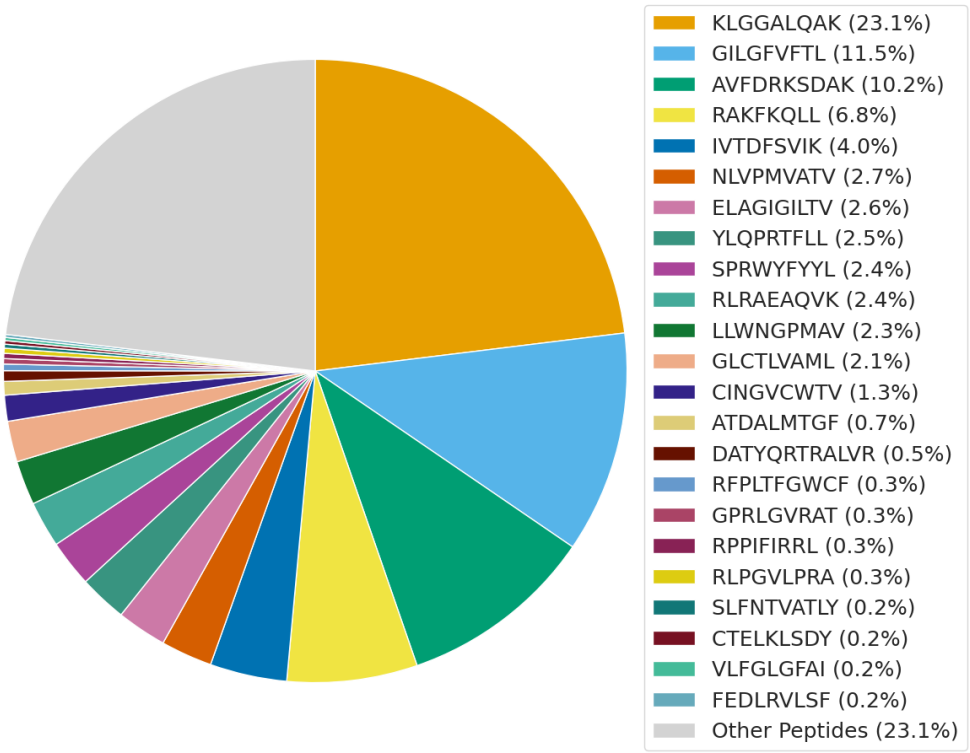

Figure 3.1 Benchmark dataset: the proportion of positive samples associated with different epitopes. The 23 epitopes given their own slices are the ones used for testing during cross validation (as described below in section 3.4.1).

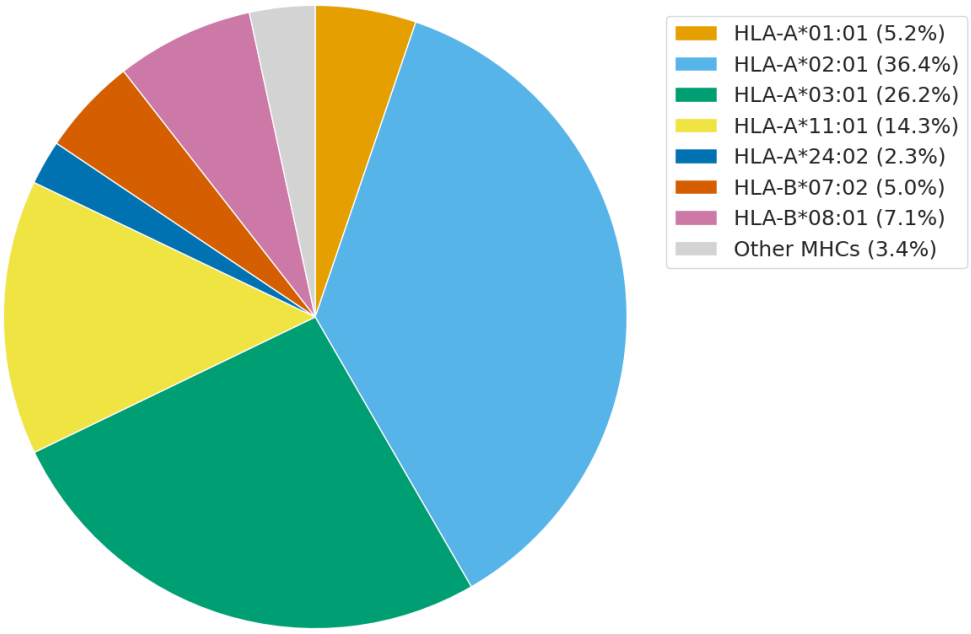

Figure 3.2 Benchmark dataset: the proportion of positive samples associated with different HLA class I alleles.

3.2.3 Negatives data samples

Most of the data sources used for the collection of positive samples (as discussed in section 3.2.1) contain few, if any, experimentally-verified negative TCR-pMHC samples (i.e. samples where there is experimental evidence that a specific TCR does not bind to a particular pMHC complex). As noted earlier (section 2.5.4), using negative data from experiments that are distinct from those used to collect the positive data, such as negative data derived from the large-scale characterisation of T cell binding within a few individuals, is liable to inflate the apparent performance of TCR-pMHC binding predictors, as they are capable of learning other distinctive features of the negative data (such as the particular combinations of HLA alleles and TCR encoding genes of the individuals from which the negative data is drawn). Consequently, for this research, negative data was generated from the positive dataset by randomly mismatching TCRs with pMHCs, a strategy widely adopted in other studies, including the recent Deng benchmark study (Deng et al., 2023). Every putative negative sample was screened against all available positive samples to exclude, as far as possible, the occurrence of randomly generated positives. Nevertheless, there remains a chance that a randomly paired TCR-pMHC will indeed

be an (as yet) unknown positive; without a better understanding of TCR cross-reactivity, the precise risk is impossible to quantify, but presumed to be small (Deng et al., 2023).

A second issue is the number of negative samples relative to the number of positives. On the one hand, the actual number of negatives (i.e. TCR-pMHC combinations where the TCR does not bind to the pMHC) that occur naturally is much greater than the number of positives. For this reason, some researchers have used datasets where the number of negatives is much higher than the number of positives, although the chosen ratio is essentially arbitrary. Examples include the ImmRep benchmarking study, which used a positive: negative ratio of 1:5 (Meysman et al., 2023).

In this research, we have chosen a 1:1 ratio, with each peptide having an equal number of positive and negative samples, an approach widely adopted in other research including the recent benchmarking paper by Deng and co-workers (Deng et al., 2023). This has the advantage that a smaller training set generally implies shorter training times – an important consideration given the constraints on access to computational resources that applied during this research. In cases (such as here) where a single peptide accounts for a large percentage of the positive samples, generating a high ratio of negative samples for that peptide becomes problematic.

3.3 Benchmark Tool Selection

3.3.1 Survey of TCR-pMHC binding prediction tools

In the past few years, more than a dozen TCR-pMHC prediction methods have been published, with an additional method – nuTCRacker (Barton et al., 2024 - unpublished) – having recently been developed in Prof Adrian Shepherd's group. In terms of the approaches and architectures employed, these methods can be broadly categorised as follows (all the nomenclature used here is explained in sections 2.2, 2.3 and 2.4):

- Decision tree-based methods: epiTCR (Pham et al., 2023), TCRex (Gielis et al., 2019) and SETE (Tong et al., 2020);
- Methods that use recurrent neural networks (including LSTMs): ERGO II (Springer et al., 2021), TCellMatch (Fischer et al., 2020) and pMTnet (Lu et al., 2021);
- Methods that use CNNs or Fully Convolutional Networks: pMTnet (Lu et al., 2021), NetTCR-2.0 and 2.1 (Montemurro et al., 2021, 2022), TITAN (Weber et al., 2021), DeepTCR (Sidhom et al., 2021), TEINet (Jiang et al., 2023) and TCRAI (Wen Zhang et al., 2021);
- BERT-based methods: TCR-BERT (K. Wu et al., 2021), TCRBert (Yoo et al., 2024) and nuTCRacker;
- Methods that use autoencoders or variational autoencoders: ERGO II (Springer et al., 2021), DeepTCR (Sidhom et al., 2021) and pMTnet (Lu et al., 2021);
- Transformer-based methods: ATM-TCR (Cai et al., 2022) , ATMTCR (Fang et al., 2022) and tcrformer (A. R. Khan et al., 2023);
- Miscellaneous other approaches: TCRDist (Dash et al., 2017) uses distance-based clustering, TITAN (Weber et al., 2021) uses a bimodal neural network architecture and TCRGP (Jokinen et al., 2021) is a Gaussian process classifier.

Note that some methods use more than one approach (e.g. pMTnet uses an autoencoder to train an embedding of TCR sequences and an LSTM to train an embedding of pMHCs) and hence appear in more than one of the above categories.

Regarding the utilisation of TCR features, most methods use only CDR3β, but others have the option of using CDR3α additionally (ERGO II) and/or alternatively (NetTCR; DeepTCR), and others require both CDR3s (TCellMatch, TCRAI). TCRGP requires a full set of CDRs (i.e. 1, 2, 2.5 and 3 for both α and β). Some methods incorporate TCR V and J gene information categorically (ERGO II, DeepTCR, TCRex, tcrformer), whereas nuTCRacker allows whole TCR sequence information to be used.

Most methods omit MHC information altogether, but some incorporate it as a 34 amino-acid pseudo-sequence (pMTnet, epiTCR, ATMTCR, nuTCRacker), whereas ERGO II incorporates it categorically. (For an explanation of HLA pseudo-sequences, see section 3.2.2). NetTCR takes a different approach: it is restricted to making predictions for the most common HLA allele, HLA-A*02:01.

Different approaches have been used to encode sequence information with BLOSUM matrices (pMTnet, epiTCR, NetTCR, TITAN, TCellMatch, TCRGP) and learned embeddings (ERGO II, TCR-BERT, TCRBert, DeepTCR, TEINet, TCRAI, ATM-TCR, ATMTCR, tcrformer, nuTCRracker) being the most popular. TCRex uses the physicochemical properties of amino acids and pMTnet uses Atchley factors (Atchley et al., 2005).

3.3.2 Selection of tools for benchmarking

The first TCR-pMHC prediction method on the list for benchmarking was the in-house tool nuTCRacker (Barton et al., 2024 - unpublished), as gaining an insight into the performance of this tool relative to others was a key motivation for this work. In choosing additional tools for benchmarking, the key criteria were that the tool had to be publicly available in a form that enabled it to be retrained using our own dataset and did not involve a lot of additional work. Tools using state-of-the-art deep learning approaches were preferred, as were those mentioned favourably in papers other than those written by the authors of a given tool.

Each model was first trained and tested on the datasets provided by the authors in order to reproduce published results (AUC values), thereby gaining confidence that the tool was working correctly, given potential differences in the versions of Python and Python libraries being used. However, a small degree of difference from published results was tolerated, as the public versions of tools are often subject to ongoing refinement leading to minor changes in the levels of performance.

What follows is a brief description of each of the six tools used in the benchmark evaluation. Note that, as was the case in the recent Deng benchmarking study (Deng et al., 2023), no attempt was made to optimise tools with respect to our benchmark dataset. No scope for optimising parameters with new data.

A) nuTCRacker

Publication: (Barton et al., 2024 - unpublished)

Architecture: DeBERTa transformer

Input features: full TCR α and β sequences, peptide, MHC represented by pseudo-sequences (see section 3.2.2); each residue treated as a token.

Commentary: hyperparameter optimisation performed on pre-trained data, with minor fine-tuning performed thereafter using a small amount of labelled data.

B) ERGO II

Publication: (Springer et al., 2021)

Architecture: LSTM, autoencoder

Input features: paired CDR3α and CDR3β sequences encoded using autoencoder; putative epitope encoded using LSTM; MHC allele encoded categorically (with, for example, HLA-A*02:01 and HLA-A*02 treated as unrelated categories); germline gene information (Vα, Vβ, Jα, and Jβ) encoded categorically.

Commentary: The following tool options were not used: the ability to make predictions for CD4+ T cells or without TCR α chain information. The ability to encode CDR3 information using an LSTM, which is a published feature of the tool, was not available in the version of ERGO II available via a GitHub repo. Non-trivial modifications to the ERGO II Python code were made to resolve library incompatibilities between those used by the ERGO II code and those installed on the Oracle Cloud Infrastructure.

C) NetTCR2.1

Publication: (Montemurro et al., 2022)

Architecture: 1D CNN

Input features: peptide and either a) all 6 CDRs (CDR123) or b) both CDR3s, with all sequences encoded using BLOSUM50.

Commentary: Both the CDR123 and CDR3 versions were evaluated.

D) TEINet

Publication: (Jiang et al., 2023)

Architecture: autoencoder, fully connected neural network

Input features: CDR3β and peptide, each encoded using a pretrained autoencoder model (TCRpeg)

Commentary: The following tool option was not used: a dynamic strategy for generating negative data.

E) ATM-TCR

Publication: (Cai et al., 2022)

Architecture: Two encoders (each with embedding layer and multi-head self-attention mechanism), linear decoder

Input features: CDR3β and peptide, each with its own encoder.

Commentary: none.

F) pMTnet

Publication: (Lu et al., 2021)

Architecture: stacked autoencoder, LSTM, deep neural network

Input features: CDR3β encoded using Atchley factors + stacked autoencoder, peptide encoded using BLOSUM50 + LSTM, MHC represented by pseudo-sequences (see section 3.2.2) and encoded using BLOSUM50 + LSTM.

Commentary: The output for pMTnet represents a percentile rank of binding strength – the smaller the rank, the stronger the predicted binding. In a pre-benchmark evaluation using the authors' own dataset, an AUC of around 0.75 was achieved (fairly close to the published AUC for this dataset) using a threshold of 0.5, with <0.5 representing a "binder". Hence a threshold of 0.5 was adopted for the subsequent benchmark evaluation.

3.4 Benchmarking strategy

3.4.1 Cross validation

To evaluate the ability of the tools described in section 3.3.2 to generalise (i.e. their ability to predict whether a given TCR will bind to peptides that are not present in the training data), a leave one group out cross-validation (LOGOCV) strategy was adopted. At a given iteration, a single group – consisting of all the positive and negative samples associated with a specific peptide – is used for testing, and all the remaining samples (none of which contain the chosen peptide) are used for training.

A "full" LOGOCV approach would entail putting each peptide (with associated samples) in its own test set for a single iteration. However, many peptides in our dataset are associated with very few TCRs. To ensure that a test set peptide has sufficient training samples to draw meaningful conclusion about tool performance with respect to that peptide, we performed partial LOGOCV; only peptides associated with at least 50 TCRs (including both positive and negative) were assigned to their own test set. There are 23 such peptides in our benchmark dataset, hence the benchmarking strategy entailed a partial LOGOCV assessment with 23 iterations.

3.4.2 Choice of benchmarking metrics

Strictly speaking, the TCR-pMHC binding prediction problem is not a binary classification task. Firstly, binding between a TCR and a pMHC complex is a matter of degree, and insofar that a meaningful binding/non-binding threshold can be specified, it is context depended (see section 1.4). Secondly, it is commonly the case that many different TCRs can bind to a given pMHC and a single TCR can bind to different pMHCs (see section 1.4).

Nevertheless, nearly all publications in this field treat the TCR-pMHC prediction problem exclusively as a binary classification task evaluated using a dataset of TCR-pMHC samples, often from diverse experimental sources, individually labelled as binders or non-binders with equal confidence; a tool makes predictions for each test set sample in turn and the area under the ROC curve (AUC) is calculated.

One important exception is the ImmRep study (Meysman et al., 2023), which calculated the epitope rank in addition to the standard AUC. For the epitope rank metric, a tool is required to make predictions for “every possible epitope seen during training”. The epitopes are then ranked according to the strength of their associated tool predictions and a score is derived from the position of the true epitope within the ranking.

It is not clear that an epitope ranking strategy can be adapted to the unseen peptide context that is the focus of this research, as the epitope of current interest will not have been seen during training. Here I have followed the standard AUC approach. One advantage is that it makes it comparatively easy to perform downstream analyses such as the correlation analyses presented in section 3.5.2. It is also worth noting that the authors of the ImmRep paper conclude that “epitope ranking mostly, but not always follows binary classification performance” (Meysman et al., 2023).

3.5 Benchmark Results

3.5.1 Generalisation performance of tools

Figure 3.3 shows the summary results for all seven TCR-pMHC binding prediction models evaluated in the benchmark. The plot shows ROC curves aggregated across all 23 epitopes selected for partial LOGOCV (as described in section 3.4.1), with a single AUC score for each tool given in the figure key. Only two methods perform notably better than random: nuTCRacker with an AUC of 0.7 and ERGO II with an AUC of 0.64.

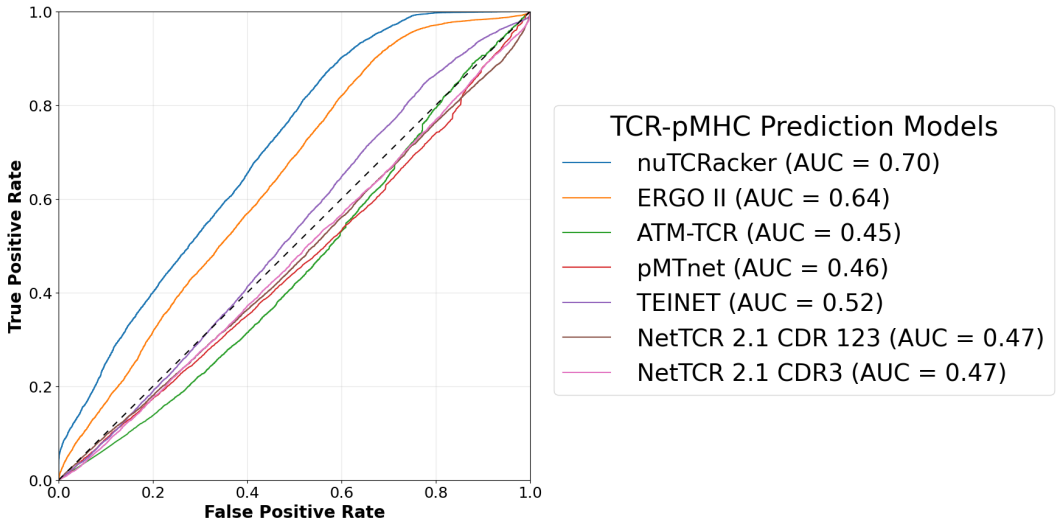

Figure 3.3 Receiver operating characteristic (ROC) curves for seven TCR-pMHC prediction models evaluated on all 23 peptides used for partial LOGOCV.

Figure 3.4 shows AUC results for the same set of tools broken down by individual peptide. Out of the 23 peptides, nuTCRacker has an AUC over 0.9 for five peptides and over 0.7 for nine peptides. The second best-performing tool, ERGO II, has an AUC over 0.9 for two peptides and over 0.7 for four peptides. The variation in predictive performance for different peptides is striking, as is the partial correlation in performance between the two methods: the four best ERGO II AUC scores are for peptides on which nuTCRacker scores very highly (AUC >0.9).

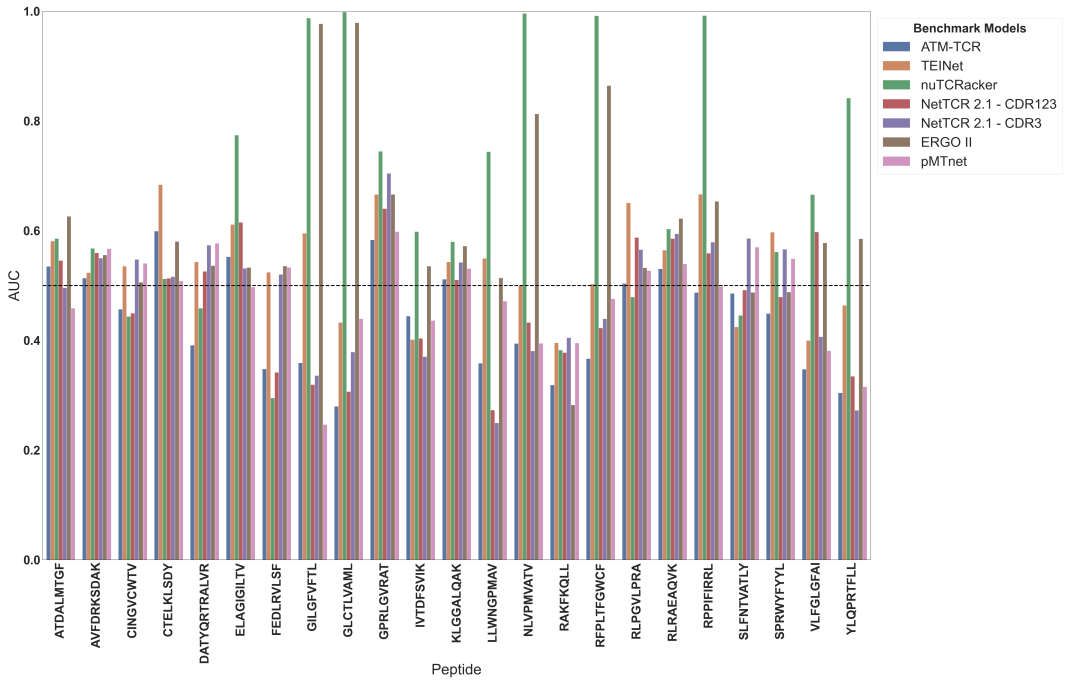

Figure 3.4 AUC scores for seven TCR-pMHC prediction models broken down by peptide. The 23 peptides used in the partial LOGOCV (see section 3.4.1) are listed on the x-axis. Each peptide has seven AUC scores associated with it represented by coloured bars, one for each prediction model. AUC is on the y-axis, with a dashed horizontal line indicating an AUC of 0.5, representing performance no better than random. The performance of the two best methods (from Figure 3.3) – nuTCRacker and ERGO II – is shown to vary greatly between peptides but to be partially correlated, with the four best ERGO II AUC scores being for peptides on which nuTCRacker scores very highly.

3.5.2 Correlation analysis

As shown in Figure 3.4, the two best performing methods – nuTCRacker and ERGO II – both score highly on a small subset of peptides. However, these peptides do not appear to have anything in common. For example, GILGFVFTL is a 9-mer binding to HLA-A*02:01 associated with over 2,000 binding TCRs in the benchmark dataset, whereas RFPLTFGWCF is a 10-mer binding to HLA-A*24:02 associated with only 61 binding TCRs, and (other than a centrally positioned phenylalanine (F)) they share no sequence-level characteristics. A third peptide, KLGGALQAK, binds to HLA-A*03:01 and is associated with over 4,000 binding TCRs. Like GILGFVFTL, it is a 9-mer with a glycine (G) at position 4. However, the AUC scores for KLGGALQAK are little better than random.

Arguably this should not come as a surprise; if there is an explanation for the different levels of predictive accuracy associated with different peptides, we should expect to find it in the relationships between the TCR-pMHC complexes in the training and test set. To gain insights into the potential factors underpinning the observed differences

in prediction accuracy for different peptides in the benchmark dataset, a range of properties were selected, each quantifying a distinct aspect of the relationship between the training and test data. The properties can be summarised as follows:

- The size of the training set (this can vary between iterations by over 20% because the number of samples associated with different test set peptides is highly variable).
- MHC frequency: the number of training set samples containing MHCs with the same HLA allele as that of the pMHC complex in the test set.
- Peptide similarity: the similarity between the test set peptide and the most similar peptide in the training set, as calculated using the BLOSUM62 score.
- TCR similarity: various properties of the TCR data were explored involving combinations of a) one more TCR sequence features (e.g. the CDR3β), with similarities calculated using TCRdist3 (Mayer-Blackwell et al., 2021), and b) the number of similar training set TCRs to be taken into account.

Having defined a set of properties, the value of a given property was calculated for all 23 peptides and the correlation between those values and the corresponding AUC scores calculated. Correlograms showing the correlations between selected properties and the corresponding AUC scores for nuTCRacker and ERGO II are shown in Figures 3.5 and 3.6 respectively.

The correlations in Figure 3.5 suggest that the ability of nuTCRacker to make accurate TCR-pMHC binding predictions for a given target peptide depends on the training set having a) a sufficient number of samples with the same HLA allele as that of the target pMHC, b) at least one peptide that is sufficiently similar to the target peptide, and c) a small number of TCRs that are sufficiently similar to those associated with the target pMHC. With respect to TCR similarity, correlations for CDR3β sequences were notably stronger than for other CDRs, which makes sense given CDR3β's pre-eminent role in forming TCR-peptide contacts. The correlations were not particularly sensitive to the number of similar training set CDR3βs that were taken into account, but 5 gave slightly higher correlation coefficients than 10 or 20. The size of the training set proved irrelevant and has not been plotted.

This is potentially useful when considering whether nuTCRacker is likely to make accurate predictions when making TCR-pMHC binding predictions involving a novel peptide, although additional guidance (How many training samples with the same HLA allele? How similar does a training peptide need to be? How similar do the TCRs

need to be?) would be much more useful. However, given that nuTCRacker gives accurate predictions (AUC >0.9) for only five peptides and “reasonable” predictions (AUC >0.7) for only nine peptides, my judgement is that an attempt to derive more detailed guidance risks overinterpretation – there is simply insufficient data. This is even clearer for ERGO II, where the correlations are weaker (Figure 3.6).

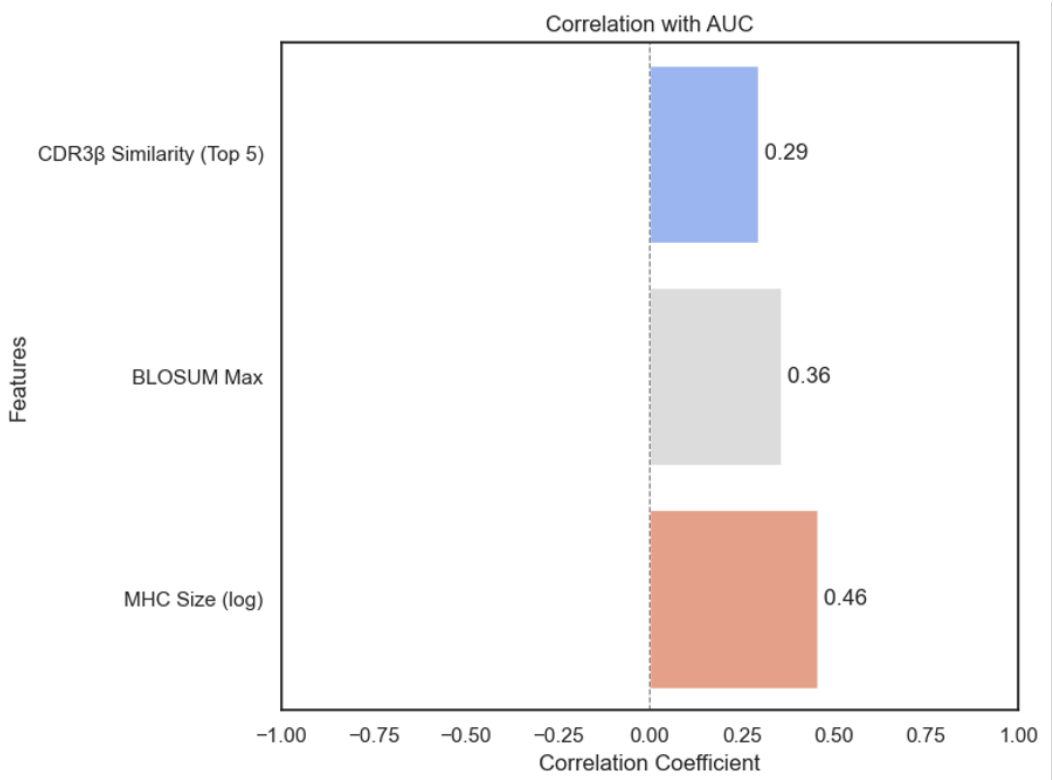

Figure 3.5 Correlogram showing the correlation between AUC and key measures of training/test relationships for nuTCRacker. CDR3b Similarity Top5= the similarity, as measured using TCRdist3, between the 5 most similar CDR3β sequences from positive samples in the training set to any (non-identical) CDR3β sequence from a positive sample in the test set; MHC Size (log) = the log of the number of training patterns having the same HLA allele as that of the test set complex; BLOSUM max = the similarity between the test set peptide and the most similar peptide in the training set calculated using the BLOSUM62 matrix. The Pearson correlation coefficient is on the x-axis. These correlations suggest that properties of the training/test set relationship associated with each of the three TCR-pMHC components contribute to overall prediction accuracy: MHC (MHC Size = 0.46), peptide (BLOSUM Max = 0.36) and TCR (CDR3b Similarity Top10= 0.29).

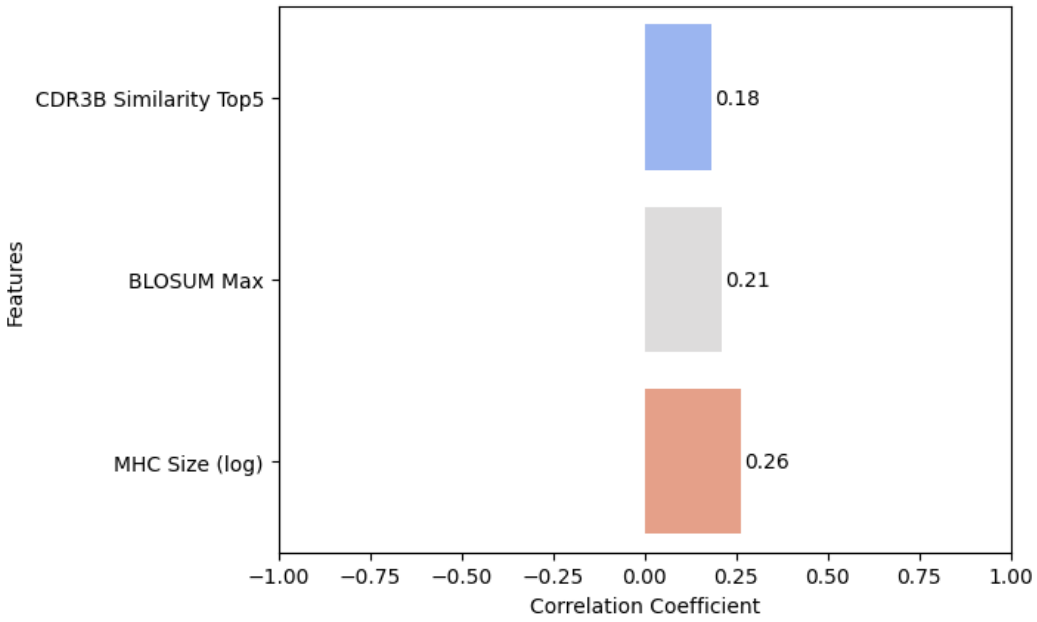

Figure 3.6 Correlogram showing the correlation between AUC and various measures of training/test relationships for ERGO II. For an explanation of the row labels, see the caption to Figure 3.5. The Pearson correlation coefficient is on the x-axis. The correlations with AUC are weaker than those for nuTCRacker (Figure 3.5), which is to be expected, given that ERGO II makes reasonable predictions (AUC >0.7) for only four of the 23 peptides.

Nevertheless, it is worth returning to the three peptides mentioned at the start of this section: GILGFVFTL and RFPLTFGWCF that score highly with both nuTCRacker and ERGO II, and KLGGALQAK that scores not much better than random. Table 3.1 shows the peptide-specific values for the three key properties plotted in the nuTCRacker correlogram (Figure 3.5) for each of these peptides, together with an additional row showing the most similar peptide in their respective training sets. Here it is clear that there are highly similar training set peptides for GILGFVFTL and RFPLTFGWCF, but no similar peptide for KLGGALQAK; this represents a plausible reason for the disparity in prediction accuracy in these cases. However, other high-scoring test set peptides do not have very similar peptide in the training set, as illustrated by two additional peptides – NLVPMVATV and GLCTLVAML – in Table 3.1.

Table 3.1 Key training/test set properties for three peptides of interest.

Peptide	**GILGFVFTL**	**RFPLTFGWCF**	**KLGGALQAK**	**NLVPMVATV**	**GLCTLVAML**
BLOSUM Max	39 (45)	62	21 (43)	27 (43)	27 (45)
MHC Size (log)	9.12	6.58	6.94	9.42	9.43
MHC	HLA-A*02:01	HLA-A*24:02	HLA-A*03:01	HLA-A*02:01	HLA-A*02:01
Same Peptide Score	45	65	43	43	45
Similarity Percent with self	86.67	95.38	58.14	67.44	60
Similar Peptide	GILEFVFTL	RYPLTFGWCF	ELAGIGALTV	VPHHVVATV	YVFCTVNAL

Note: In the final row, amino acid residues that match the target peptide at a given position are show in red.

One factor that is shared by six of the nine peptides for which nuTCRacker has AUCs over 0.7 (including NLVPMVATV and GLCTLVAML) is their association with HLA-A*02:01, the most common HLA allele in the dataset. (An additional four, poorly predicted test peptides are also associated with HLA-A*02:01.)

3.6 Discussion

In the benchmarking study presented in this chapter, two tools showed a moderate ability to generalise to unseen peptides, with potentially useful levels of prediction accuracy (AUC >0.75) being achieved by nuTCRacker in around a third of cases, but less than a quarter of cases with ERGO II. This outcome challenges the widespread assumption that even "modern deep learning methods" cannot generalise (to quote (Grazioli et al., 2022)); in a non-trivial proportion of cases, generalisation is possible. Moreover, although two of the most accurately predicted test set peptides had, in their corresponding training sets, peptides that differ by only a single residue, other well-predicted peptides had no close training set peptides (see NLVMVATV and GLCTLVAML in Table 3.1), demonstrating that this is not merely a case of marginal levels of generalisation to near-identical samples.

Although nuTCRacker employs state-of-the-art deep learning techniques and utilises more TCR features than nearly every other method, this is not true for ERGO II. Explaining why these two tools performed better than the others has proved elusive.

The correlation analysis in section 3.5.2 suggests that all three molecular components — TCR, peptide and MHC — made contributions to prediction accuracy, and that having a training set that contains a few similar TCRs, at least one similar peptide and a good number of samples with the same HLA allele is an indication that predictions have the potential to be accurate. However, even from the small number of contrasting examples shown in Table 3.1, it seems clear that different combinations of factors are contributing when accurate predictions are made for different unseen peptides. It is likely that there are additional factors that have yet to be explored that may provide further insights, for example, the relationships between different HLA alleles (e.g. in terms of HLA supertypes (Sidney et al., 2008)), or biases within the dataset (e.g. with respect to the number of peptides associated with each HLA allele, a topic addressed in section 4.4.4 in a different

context). However, it is likely that the various factors are connected in complex ways, with the impact on prediction accuracy varying in different parts of TCR-pMHC space; it seems unlikely that there are simple and universal insights that have yet to be discovered.

In terms of different ways of approaching the research presented here, the BLOSUM62 matrix is unlikely to be the optimal way of measuring peptide similarity, as it is designed to score alignments between sequences that are evolutionarily divergent. A better approach would consider what constraints should be applied to the alignment of antigenic peptides, given knowledge about docking footprints and how peptides of different length are accommodated in the MHC binding groove.

Finally, given more computational resources, it would be worth exploring the impact of using different datasets and different ways of handling the binding/non-binding threshold. The “optimal" dataset will be tool dependent, but also vary between applications. For example, a tool/dataset combination that performs well when ranking the likelihood of TCRs binding to a known pMHC may be poor at repertoire annotation.

4 Developing a Tool to Predict TCR-HLA Associations

4.1 Introduction

The TCR-pMHC prediction task addressed in Chapter 3 makes the implicit assumption that the target MHC complexes in the context of the peptide that it presents to the TCR, are known in advance – the question is: Which of the available TCRs does a given pMHC bind to? However, a common scenario arising from single-cell repertoire sequencing initiatives is that the TCRs in the repertoire is known, but not the pMHCs.

In such circumstances, it is worth getting a feel for the vastness of peptide-MHC space that one would theoretically need to explore in order to pin down the pMHC target of every TCR in a repertoire. Regarding the likely peptide targets, a T cell repertoire will contain a diverse population of long-lived memory cells targeting specific peptides within the pathogens that the host has encountered, and most of those pathogens are unlikely to have been identified. For example, it is estimated that adults catch two or three common colds a year on average, and over 200 viral strains have been identified that cause cold symptoms, some of the commonest being from widely different regions of the viral taxonomic classification (e.g. rhinoviruses and adenoviruses (Mackie, 2003)). With regards to MHC class I, there are currently nearly 27,000 HLA-A, -B and -C alleles in the IPD-IMGT/HLA Database (Barker et al., 2023) (URL www.ebi.ac.uk/ipd/imgt/hla/index.html, accessed 07/09/2024) and there may be millions yet to be discovered[8]. Bearing these points in mind, it is easy to see that a strategy capable of reducing the vast combinatorial challenges posed by unconstrained peptide-MHC space are likely to be of practical benefit when making TCR-pMHC binding predictions in the context of TCR repertoires. One such strategy is to predict the likely HLA restriction of a given TCR.

[8] James Robinson, Senior Bioinformatics Scientist at the Anthony Nolan Research Institute, refers to an estimated total of 8 to 9 million HLA-I alleles in an Anthony Nolan blog posting at https://www.anthonynolan.org/blog/2017/08/hla-typing-our-lifesaving-research

The focus of this chapter is on the development of a tool for predicting the likely MHC target of a human CD8+ TCR irrespective of the peptide bound to that MHC molecule. This task is formulated in terms of TCR and MHC in isolation by, in effect, removing the peptides from TCR-pMHC complexes. Given that TCRs do not bind to MHC molecules without a peptide being present, I describe this as the task of predicting TCR-HLA associations rather than TCR-MHC binding, although both terms are used in the literature (see section 4.1.2). If a TCR is said to "have an association" with a particular type of MHC molecule, it implies that the TCR can form a TCR-pMHC complex involving that MHC provided the latter has a suitable peptide bound within its groove.

In the remainder of this section, I will provide a more detailed introduction to the task of predicting TCR-HLA associations (section 4.1.1) and discuss other attempts to address it (section 4.1.2), before moving on to consider: the creation of a dataset for evaluating a TCR-HLA associations prediction method (section 4.2); explaining the tool I built to address the task (section 4.3); the choice of evaluation strategies and performance metrics (section 4.4); and the performance of the tool on both "seen" and "unseen" HLA types, together with a correlation analysis exploring the relationship between properties of the data and predictive performance (section 4.5).

4.1.1 On the prediction of TCR-HLA associations

To gain a sense of the nature and scale of the challenge involved in the prediction of TCR-HLA associations, it is worth considering important aspects of the underlying biology that shape this task.

When a TCR binds to a pMHC, the number and location of the contacts between the TCR and the MHC molecule are very different to those between the TCR and the peptide. In brief, the MHC residues that are contact with the TCR are mainly along the exposed "top" of the α-domains that form the MHC groove, and multiple CDRs are likely to be in contact with the MHC molecule, although the position of the TCR with respect to the MHC's α-domains and which combination of CDRs are involved can vary considerably, depending on the orientation of the two molecules (for a more detailed discussion, see section 1.4.2). Whereas many TCR-pMHC prediction tools focus exclusively on the sequences of the CDR3β and peptide, this is clearly

inappropriate from the perspective of TCR-HLA associations. The choice of TCR and MHC features for addressing this task is discussed further in section 4.2.

TCRs are, to a significant extent, HLA restricted, i.e. there is a strong tendency for TCRs to bind exclusively (or predominantly) to a single type of MHC molecule (or closely related MHC molecules). The wording in the preceding sentence reflects the fact that the extent of HLA restriction is hard to pin down, as a particular TCR may only bind to a given MHC molecule when the latter is in complex with one or more specific peptides (and there is a near-infinite number of such peptides), and most MHC molecules are, as noted earlier, yet to be discovered. Hence, strictly speaking, the TCR-HLA associations prediction task (much like the TCR-pMHC binding prediction task, as discussed in section 3.4.2) is not a strictly binary classification task. However, here I have chosen to treat it as a binary classification task for two main reasons: consistency with previous work on this problem (see section 4.1.2); and to simplify the correlation analyses presented in section 4.5.

When the topic of TCR-HLA associations is considered from the perspective of TCR repertoire annotation, potentially useful constraints may apply. It is reasonable to assume that all the TCRs present within a single repertoire will have an association with at least one of the MHC molecules expressed by the individual from whom the repertoire was collected, given how the T cell thymic selection process works (see section 1.1.3). That individual will express MHCs associated with a limited number of HLA alleles – one allele (in the homozygous case) or two alleles (in the heterozygous case) — for each of the different types of HLA gene, such as HLA-A and HLA-B (MHC class I genes are discussed briefly in section 1.2.1). When repertoires come from HLA-typed individuals, this limited number of HLA alleles will be known. For individuals that have not been HLA-typed, if their ethnic origin is known, the most common HLA alleles observed in that ethnic group can be ascertained (at least for ethnicities that have been well studied) from the Allele Frequency Net Database (Gonzalez-Galarza et al., 2020) (www.allelefrequencies.net/hla.asp).

Regarding the target antigen, in cases where pre- and post-vaccination repertoires[9] are available, clonal expansion observed in a post-vaccination sample is often

[9] Although a majority of the vaccines are antibody based, there can be an internal CD8+ response and that can reduce the risk of disease caused by pandemic viruses (Gilbert, 2013).

assumed to have arisen in response to antigens belonging to the pathogen that was targeted by the vaccine (although a response to other components of the vaccine, or to an unrelated infectious agent, is not uncommon). More generally, it is known that individuals who share the same HLA allele may target the same "public" epitope, and some of these public epitopes have already been identified. For example, a recent study found that all 50 members of a cohort of COVID-19 convalescents sharing at least one common HLA-I allele produced T cells that recognised four SARS-CoV-2 epitopes, with an additional four epitopes recognised by 80% or more of that cohort (Meyer et al., 2023).

Ultimately, the most useful and generally applicable constraint is arguably the fact that TCRs from the repertoire of a single individual will have associations with only a small number of different MHC molecules – only six "classical" MHC-I molecules if the individual is heterozygous with respect to HLA-A, HLA-B and HLA-C – and that number is fixed (in contrast to the vast and changing "population" of antigenic peptides that TCRs may encounter). Given a sufficiently accurate prediction method, it may be feasible to work out an individual's HLA type computationally. For example, if the number of predicted TCR-HLA associations is notably higher for two HLA-A alleles, the individual is likely to be heterozygous at this locus and any predicted associations involving other HLA-A alleles are likely incorrect. However, in cases where the number of predicted TCR-HLA associations is notably higher for only a single HLA-A allele, it may be infeasible to distinguish between a homozygous individual and one who has a second HLA-A allele that (as is not uncommon) is expressed at low levels (for an investigation of differential expression levels of HLA-A alleles, see (Ramsuran et al., 2015)).

4.1.2 Previous work on predicting TCR-HLA associations

Existing methods that predict TCR-HLA associations can be broadly split into three categories: HLA-typing methods, antigen specificity-based methods, and general TCR-HLA association methods.

HLA-typing methods focus on predicting the HLA-type of the individual from which a TCR repertoire was derived and are deemed to have succeeded if the true HLA-type of the individual is predicted correctly irrespective of the number (or proportion) of TCR-HLA associations correctly identified. An example is the HLA-typing method

developed by Ortega and co-workers, who undertook a large-scale statistical analysis of HLA-typed datasets to identify TCRα and TCRβ chains likely to be associated with specific HLA alleles, and subsequently created HLA-specific logistic regression models to predict whether a given repertoire is associated with particular alleles. The authors report an AUC of 0.96 for their TCR β chain models when applied to an independent dataset of 46 HLA-typed individuals (Ortega et al., 2024). This work overlaps with other research into the occurrence patterns of public TCRs (i.e. TCRs found in multiple individuals) – see, for example, the paper "Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity" (DeWitt et al., 2018).

Methods that focus on the identification of TCR antigen specificity may address HLA restriction as a secondary consideration, based on the observation that TCRs binding to the same antigen and sharing similar sequences commonly also bind to the same MHC molecule. Typically, such methods focus on the identification of conserved CDR3 motifs. For example, Glanville and co-workers identified five CD4+ TCR antigen specificity groups in 22 *Mycobacterium tuberculosis*-infected individuals of which four were associated with one or two HLA-II alleles (Glanville et al., 2017).

The research presented in this thesis fits into the third category of method – those that seek to predict TCR-HLA associations without directly considering whether a given TCR is public or has a particular antigenic target. I am aware of two methods having been published so far: CLAIRE (Glazer et al., 2022) and DePTH (Liu et al., 2023). Both are worth considering in some detail.

CLAIRE (Glazer et al., 2022) is a deep learning predictor that uses a pre-trained autoencoder and CNN with two convolutional layers to predict CD4+ and CD8+ TCR-HLA associations given the following input: information about the TCR β chain and optional α chain consisting of the CDR3 sequence plus V and J gene information encoded categorically; and the HLA allele encoded categorically. Different versions of CLAIRE were developed using different sets of data; the version tested using McPAS data achieved an AUC of 0.87, whereas the version tested using VDJdb

achieved an AUC of 0.72[10]. Additional tests that involved applying the different versions of CLAIRE to multiple bulk sequence datasets were less successful, with nearly all AUCs less than 0.6 and several at or close to 0.5.

DePTH (Liu et al., 2023) is a neural network-based predictor with multiple fully connected layers and a separate CNN for handling CDR3 data only. Like CLAIRE, it makes predictions for both CD4+ and CD8+ TCR-HLA associations and works with the following input: sequences for all six standard CDRs plus CDR2.5 and HLA allele pseudo-sequences. DePTH was trained and tested using public TCRs (i.e. TCRs found in a minimum of two individuals) from the repertoires of over 600 HLA-typed individuals in the CMV study conducted by Emerson and co-workers (Emerson et al., 2017), with the HLA associations of the public TCRs determined by a preliminary statistically assessment. 6,323 TCR-HLA-I pairs were identified with five times as many negatives as positives (the relative HLA frequencies are the same for positives and negatives). The authors report an AUC of 0.82 for TCR-HLA associations when the HLA is also present in the training set, falling to AUCs of between 0.64 and 0.69 in "unseen" HLA mode, with DePTH successively retrained with TCR-HLA associations for the three most common HLA-I alleles (HLA-B*08:01, HLA-B*07:02 and HLA-C*07:01) omitted from the training set and used exclusively for testing.

4.2 TCR-HLA Associations Datasets

This section focuses on the creation of datasets suitable for evaluating the tool I developed for predicting TCR-HLA associations (the tool itself is described below in section 4.3).

4.2.1 Labelled data sources, filtering and negative data

For this task, the TCR-pMHC binding prediction benchmark dataset described in section 3.2 – containing data drawn from five public resources: VDJdb (Shugay et

[10] After multiple re-readings of the CLAIRE paper, I remain uncertain what dataset (or datasets) were used for training CLAIRE with respect to these two results. On the one hand, it is clear that there is a version of CLAIRE trained exclusively on McPAS; on the other hand, it is stated that "We used again the McPAS, and enlarged it with the VDJDB dataset". The latter is shortly before the following: "The AUC obtained from the McPAS dataset was significantly better than the one from VDJdb (AUC 0.87 versus 0.72; Fig 4)."

al., 2018), McPAS-TCR (Tickotsky et al., 2017), TBAdb (Wei Zhang et al., 2020), IEDB (Vita et al., 2019), and 10x Genomics public datasets (10X Genomics, Pleasanton, CA, USA; www.10xgenomics.com/datasets/) – was repurposed for the prediction of TCR-HLA associations. Two major changes were applied to this original dataset.

Firstly, peptides were removed from all records. Given that a small proportion [602 out of 32,000] of the original records contained the same combination of TCR and MHC bound to different peptides, this produced a corresponding number of duplicate records, which were removed. In 12 cases, two or more TCR-HLA associations were found to contain the same TCR bound to different HLAs. Given the decision to treat the prediction of TCR-HLA associations as a binary problem, it was decided that, in these cases, only one, randomly selected TCR-HLA association would be retained in the dataset. (In retrospect, a cleaner solution would have been to simply remove all such records.)

A key motivation in this research was to investigate HLA-specific correlations (analogous to the analysis of TCR-pMHC binding predictions presented in section 3.5.2). To ensure that all dataset HLAs have sufficient training samples to draw meaningful conclusion about tool performance, only HLAs associated with at least 40 TCRs in positive samples (implying a minimum of 80 records per HLA when negatives are taken into account) were retained. Within the remaining dataset, more than 50% of the samples were associated with a single HLA (HLA-A*03:01). These were downsampled to bring the number of TCR-HLA-A*03:01 associations close to the number for the second most highly represented HLA allele (HLA-A*02:01), leaving a total of 18,580 positive records.

In the absence of true negatives, the equivalent strategy to that adopted for the TCR-pMHC benchmark dataset (see section 3.2.3) was adopted here, with the same number of negative records being generated for a given HLA allele as the number of positive records associated with that allele, thereby double the total number of records to 37,160. After removing the duplicates found in both positive and negative records, the final dataset contained 35,890 records. This is what I will call the *primary dataset*. The final proportion of records associated with different HLA alleles in the primary dataset is shown in Figure 4.1. This dataset was used for the preliminary evaluation of DeepTHAWT's performance to predict the association between a given

TCR and HLA using a standard 10-fold cross validation strategy as described in 4.4.1.

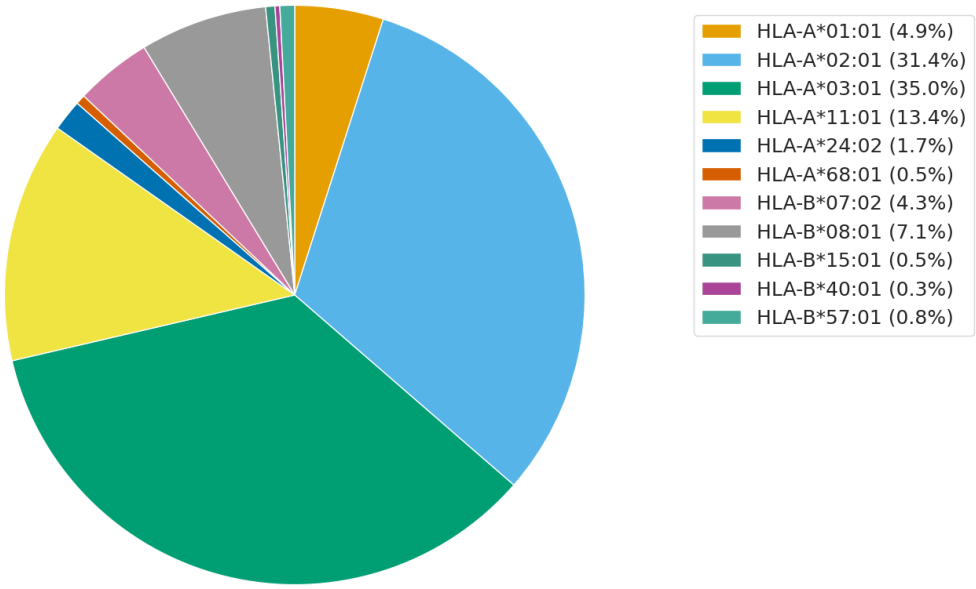

Figure 4.1 TCR-HLA associations in the primary dataset. The proportion of records associated with different HLA class I alleles.

A second evaluation of the tool I developed was performed using non-overlapping subsets of the data from the primary dataset, namely VDJdb data for training and 10x Genomics data for testing. In both cases, HLA-A*03:01 was the most common HLA allele and its records were downsampled such that the number of HLA-A*03:01 records was close to the number of the second most common HLA allele in both datasets, HLA-A*02:01. This left a total of 14,319 positive samples in the VDJdb training set and 2,979 positive samples in the 10x Genomics test set, with an equal number of negatives per HLA in both sets. (Note that the combined total of positive records for these two datasets is less than 1,000 records fewer than the total for the whole primary dataset. This reflects the comparatively small number of unique records available from the other sources that contributed to the primary dataset.)

4.2.2 Selection of features based on biological knowledge

As discussed in section 1.4.2, all six “classical” CDRs of a TCR are potentially involved in forming contacts with the MHC molecule, as well as the additional variable loop located between CDRs 2 and 3 known as CDR2.5 (Dash et al., 2017). The sequences of all eight of these CDRs are used as inputs to the TCR-HLA associations prediction tool presented in this thesis. The is the same set of TCR features chosen by the authors of DePTH (Liu et al., 2023) but differs from the set used by CLAIRE (Glazer et al., 2022), for which only CDR3 sequences encoded (with V and J gene information encoded categorically).

In the TCR-pMHC binding prediction benchmark dataset, an MHC molecule is represented by a pseudo-sequence comprising a set of 34 residues that are putatively in contact with the peptide (see section 3.2.2). The same choice is made by the authors of DePTH (whereas CLAIRE encodes HLA allele names categorically). However, whereas pseudo-sequences appear a good choice for the TCR-pMHC binding prediction task, direct contacts between TCR and MHC potentially involve a different and larger set of MHC residues, as discussed in section 1.4.2. For these reasons, the entire sequences of the MHC α1 and α2 domains were selected for the model developed in this research.

4.2.3 Unlabelled TCR data

A key feature of the tool I developed is that it incorporates a pre-trained model of unlabelled TCRs, an approach known as transfer learning. (Transfer learning is discussed in section 2.4; details of the approach adopted here are given below in section 4.3.) To train this model, bulk TCR data was downloaded from two sources: iReceptor (Corrie et al., 2018) and 10x Genomics (10X Genomics, Pleasanton, CA, USA; www.10xgenomics.com/datasets/). In both cases, data was selected to include only productive sequences from human loci TRA or TRB having the CDR3, and V and J genes completely specified, with an additional “high confidence” criterion applied when downloading data from 10x Genomics. After the removal of duplicates, there remained a total of 410,499 TCRs with paired α + β chains and 24,429,131 unpaired chains from iReceptor, and 202,094 paired records from 10x Genomics. MHC Class I sequences were downloaded from MHC Motif Atlas at http://mhcmotifatlas.org/class1 (Figure 4.2).

All labelled and unlabelled data was ultimately converted into numerical vectors using the Hugging Face Transformers Tokenizer (see section 2.4.3; further details are given in Appendix 1).

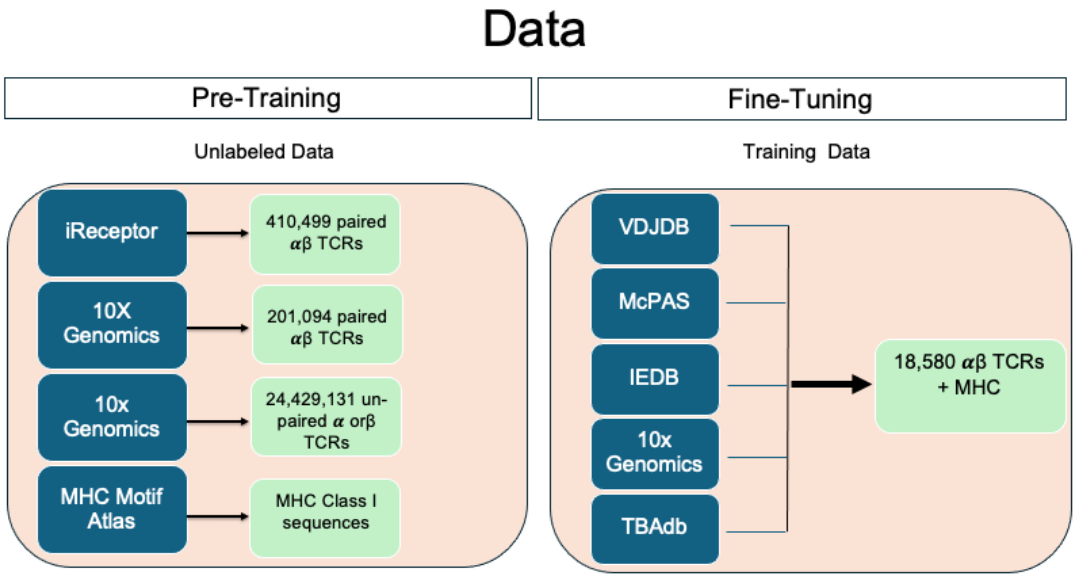

Figure 4.2 Schematic Overview for DeepTHAWT Data. The pretraining TCR sequence data was collected from iReceptor and 10X Genomics. The MHC sequences were collected from MHC motif atlas. The fine tuning paired TCR-MHC data was collected from public repositories.

4.3 Architectures and Algorithms

The DeepTHAWT (**Deep** learning **T**CR-**H**LA **A**ssociations predictor **W**ith **T**ransformer) tool developed to tackle the prediction of TCR-HLA associations implements a transfer learning approach (discussed in section 2.4). The most important component is the DeBERTa (Decoding-Enhanced BERT with Disentangled Attention) architecture (He et al., 2020). A DeBERTa-based transformer model was pre-trained using unlabelled TCR data (described above in section 4.2.3) and unlabelled MHC data (see below). The training was performed using masked language modelling (see section 2.4.4) by randomly masking 40% of the residues.

Having developed the pre-trained model, it was available for fine-tuning (i.e. undertaking additional training) using a dataset of labelled TCR-HLA associations

(see section 4.2.1). At this stage, there is an additional requirement – that the associations data is classified (in this instance a binary classification, i.e. "Is this record a real TCR-HLA association?", true or false, 1 or 0). This is handled using a simple neural layer, known as a classification head, with the final binary prediction made via the application of the SoftMax function. The DeepTHAWT architecture is summarised in Figure 4.3.

The choice of data features incorporated in the final model was based on a combination of prior knowledge about the contacts between TCRs and MHC molecules and subsequent computational evaluations. In a preliminary assessment at the pre-training stage, using full TCRs sequences was deemed impractical given available GPU resources, so TCR representation was limited to the eight "non-classical" CDRs (including both CDR2.5s). In a preliminary evaluation at the fine-tuning stage, representing MHCs by the complete sequences of their α1 and α2 domains (downloaded from MHC Motif Atlas at http://mhcmotifatlas.org/class1) proved superior to using HLA pseudo-sequences (as described in section 3.2.2 and used by DePTH (Liu et al., 2023)). (The final pre-trained model has been saved in Shepherd Group's private Hugging Face repository.)

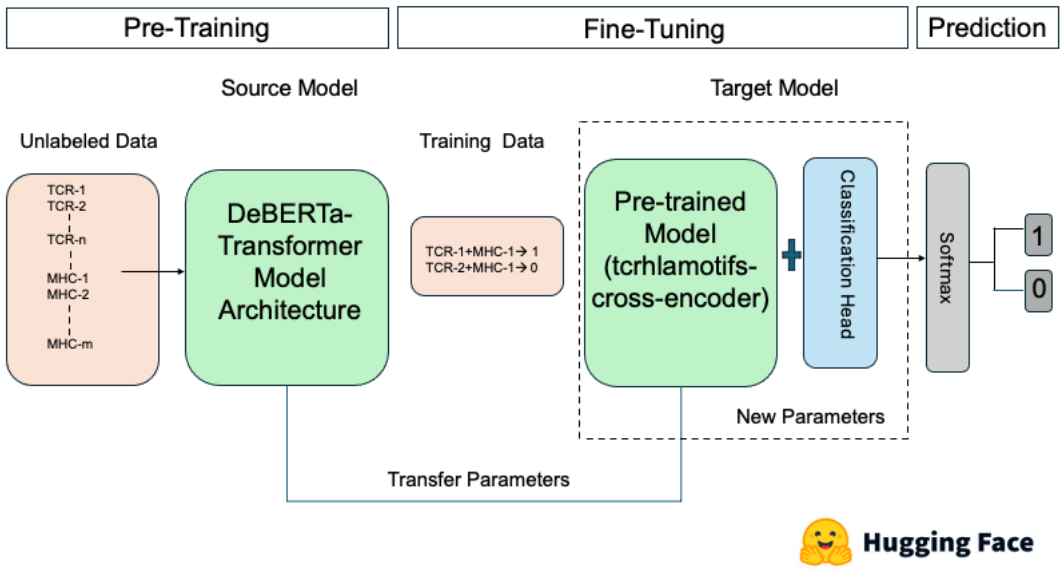

Figure 4.3 Schematic overview of the DeepTHAWT architecture. The model is developed using an open source machine learning framework called PyTorch[11] which is based on Python and the Torch library. It uses the DeBERTa architecture and is implemented with Hugging Face Transformer Architecture.

[11] https://pytorch.org/

4.4 Results for the Prediction of TCR-HLA Associations using DeepTHAWT

4.4.1 DeepTHAWT 10-fold cross validation results

A preliminary evaluation of DeepTHAWT's performance was undertaken using a standard 10-fold cross-validation strategy on the primary dataset (described in section 4.2.1). In this evaluation, no attempt was made to prevent HLAs in the test sets from being present in the training set – i.e. the evaluation assesses the performance of the tool on "seen" HLAs – although no test set TCR was ever present in the corresponding training set.

DeepTHAWT achieved an overall AUC of 0.68. This is considerably lower than the AUC of 0.82 reported for DePTH (Liu et al., 2023) and the AUC of 0.87 reported for CLAIRE on the McPAS dataset (Glazer et al., 2022), but quite close to the 0.72 achieved by CLAIRE using VDJdb data. As in the case of TCR-pMHC binding prediction tools (that pre-empted the benchmarking research of Chapter 3), differences in the datasets used for training and testing are likely to be an important contributory factor here; indeed, the multiple AUC scores for different datasets reported in the CLAIRE paper demonstrate this point (Glazer et al., 2022).

DeepTHAWT's performance broken down by individual HLAs is shown in Figure 4.4. The highest AUC score is for HLA-A*02:01, which accounts for 31.4% of the data, whereas the lowest scores are for HLA-A*68:01 and HLA-B*40:01, which account for 0.5% and 0.3% of the data respectively (see Figure 4.1). However, the most highly represented HLA in the dataset, HLA-A*03:01 (35% of the data), has an AUC that might reasonably be described as "average". Potential explanations for this disparity together with an exploration of factors that may contribute to the performance differences observed with different datasets is deferred until section 4.4.4.

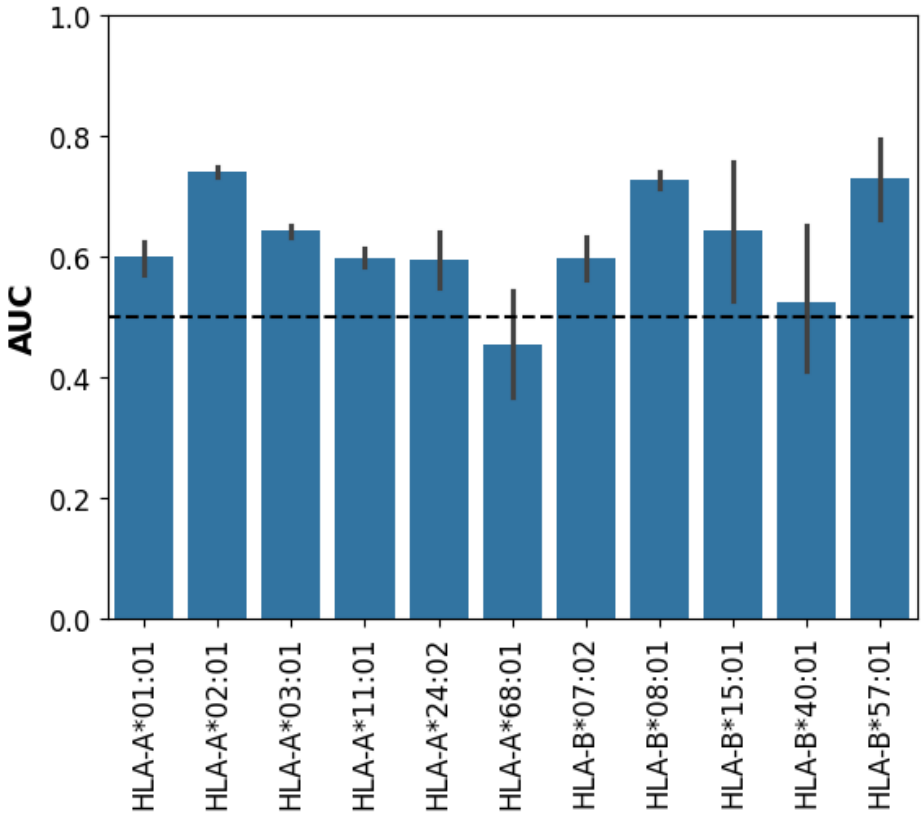

Figure 4.4 DeepTHAWT AUC scores for the primary dataset (aggregated across the 10 cross-validation folds) broken down by HLA. Each bar shows the AUC calculated from all the predictions involving a given HLA (i.e. *not* by averaging the per-fold AUCs). Each error bar shows the variation in per-fold AUC scores for a given HLA. The three HLAs with the largest error bars – HLA-A*68:01, HLA-B*40:01 and HLA-B*15:01 – have the smallest number of records in the dataset, of which the first two also have the lowest AUC scores. Proportion of records associated with each HLA are shown in figure 4.1.

4.4.2 DeepTHAWT trained and tested on different datasets

The primary dataset used in the preceding evaluation (section 4.4.1) is highly diverse, combining data from five distinct sources: McPAS, VDJdb, TBAdb, IEDB and 10x Genomics (see section 3.2.1). To gain a different perspective on DeepTHAWT's performance, it was retrained using only the data from VDJdb – the largest contributing source to the primary dataset – and tested on the 10x Genomics data. As noted in section 4.2.1, HLA-A*03:01 (the most common HLA allele in both datasets) was downsampled in both the training and test sets. The final train and test datasets were checked for overlap, and if any were found, those records were removed from the train set.

With this combination used for training and testing, DeepTHAWT achieved a higher AUC score of 0.76. This is much closer to the AUC scores achieved with CLAIRE (with McPAS and VDJdb) and DePTH (on a set of public TCRs). DeepTHAWT's

performance on the four HLAs with sufficient data in the 10x Genomics dataset is shown in Figure 4.5. These results will be considered with those for the primary dataset (section 4.4.1) in section 4.4.4 below.

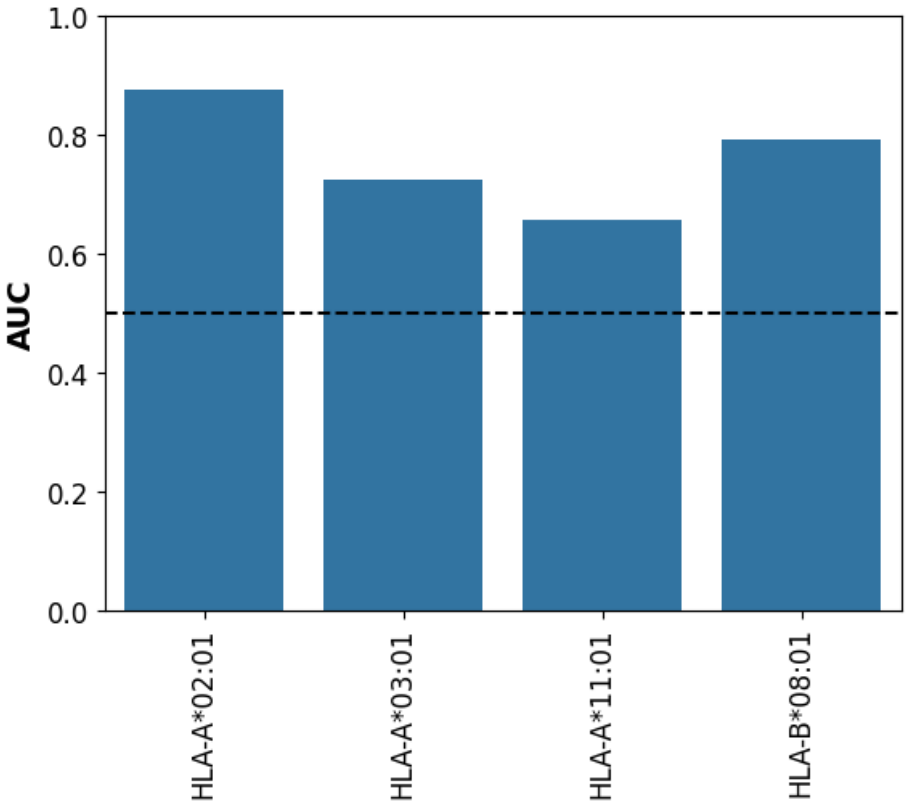

Figure 4.5 HLA-specific AUC scores for DeepTHAWT trained on the VDJdb dataset and tested on the 10x Genomics dataset. There is no correlation between predictive performance and the number of per-HLA records in either the training or test sets.

4.4.3 Generalisation performance of DeepTHAWT

To test DeepTHAWT's ability to generalisation to unseen HLAs, an equivalent leave one group out cross-validation (LOGOCV) strategy was adopted as previously used in the evaluation of TCR-pMHC binding prediction tools (sections 3.4.1 and 3.5.1). Here, the primary dataset (see section 4.2.1) was used for the evaluation; at each iteration, the TCR-HLA associations involving a single HLA allele were omitted from the training set and used exclusively for testing.

DeepTHAWT achieved an overall AUC of 0.50, in other words no better than random. This compares with AUCs between 0.64 and 0.69 when DePTH was applied to the three most common HLA alleles in its dataset (Liu et al., 2023). The performance of DeepTHAWT on individual HLAs is shown in Figure 4.6. The

performance on HLA-A*02:01, the second most common HLA allele in the dataset, is particularly poor. This will be addressed in the next section.

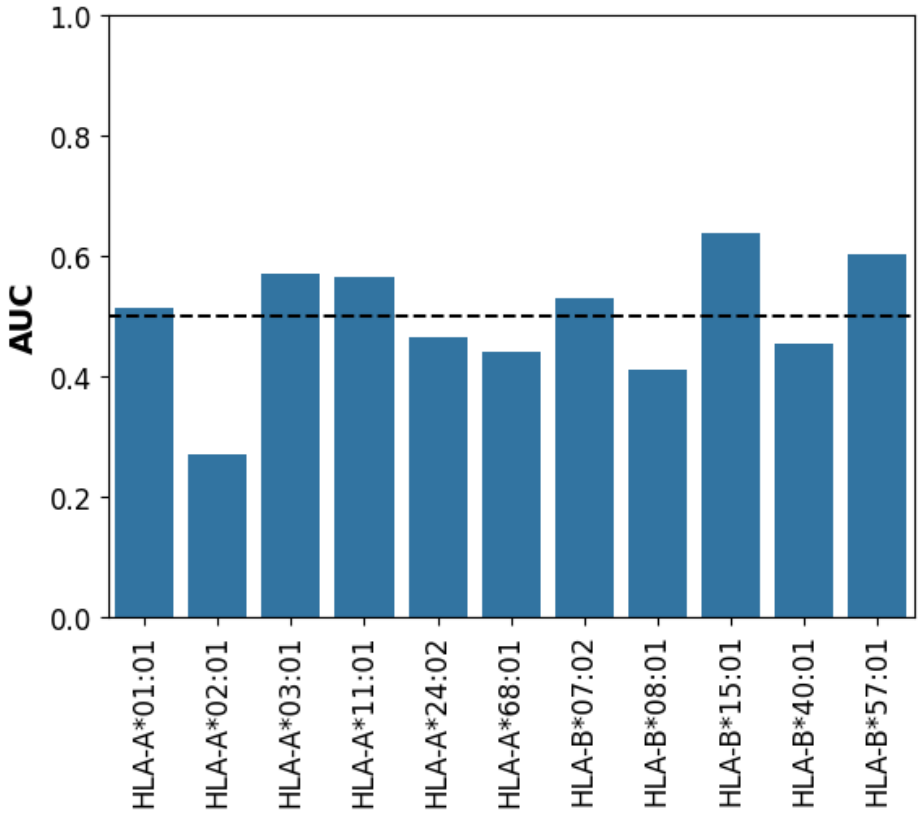

Figure 4.6 Generalised HLA-specific AUC scores for DeepTHAWT trained on the primary dataset (LOGOCV strategy). The proximity of most AUC scores to 0.5 reflects the overall, poor generalisation performance of the tool, with HLA-A*02:01 (the second most common HLA allele in the dataset) a striking and surprising worse-than-random outlier.

4.4.4 Dataset analysis

In this section, I will consider the underlying properties of the TCR-HLA associations data that may help to explain the disparities in performance observed between CLAIRE (Glazer et al., 2022), DePTH (Liu et al., 2023) and DeepTHAWT for different datasets and HLA alleles.

DePTH stands out from the other two methods because it was trained and tested exclusively using public TCRs (Liu et al., 2023). It is worth noting that public TCRs may have characteristics that make them somewhat easier targets for prediction than TCRs in general. For example, it has been observed that public TCRs tend to be close to germline with the addition of few nucleotides during the process of V(D)J recombination (Venturi et al., 2006); V(D)J recombination is described briefly in section 1.3.1).

Regarding the data used by CLAIRE (Glazer et al., 2022) and DeepTHAWT, there is clearly a non-trivial degree of overlap between their respective datasets, as both incorporate data from McPAS and VDJdb. However, DeepTHAWT's more stringent requirements (paired, full-length TCR α and β chains compared to CLAIRE's CDR3 sequence plus V and J gene annotations with α chain optional) mean that the quantity of suitable records is vastly different for the two tools. Hence, it is reported that CLAIRE's VDJdb dataset contained 40,000 distinct TCRs (Glazer et al., 2022), whereas the total number of distinct TCRs in the DeepTHAWT primary dataset, drawn from five public sources (including VDJdb), is only 18,580 (and – at least below a certain, relatively high threshold – additional training data often leads to improved performance with machine learning prediction tools).

The disparity in the number of primary dataset samples (equivalent to the number of TCRs) per HLA allele has already been summarised in Figure 4.1. In what follows, it is useful to bear in mind that, taken together, alleles HLA-A*02:01 and HLA-A*03:01 account for over 65% of the records in the primary dataset, each with over 30% of the total number. Beyond the disparity of TCRs per HLA allele, there is a further disparity in terms of the number of unique peptides associated with different alleles. More than half of the unique peptides (well over 400) from the original set of TCR-pMHC complexes (i.e. prior to the removal of the peptides and subsequent removal of duplicates) bind to MHC molecules that are encoded by a single HLA allele, HLA-A*02:01. The distribution of unique peptides between different HLA alleles is shown in Figure 4.7. Contrastingly, allele HLA-A*03:01 – associated with the largest number of TCRs in the dataset – is only associated with 13 peptides from the original TCR-pMHC set, and nearly all the HLA-A*03:01-associated TCRs bind in the presence of a single peptide, as shown in Figure 4.8. Although peptides are not represented in the TCR-HLA association prediction task, peptide diversity is linked to TCR diversity – specifically, TCRs binding to the same peptide are more likely to be similar to each other than TCRs binding to different peptides.

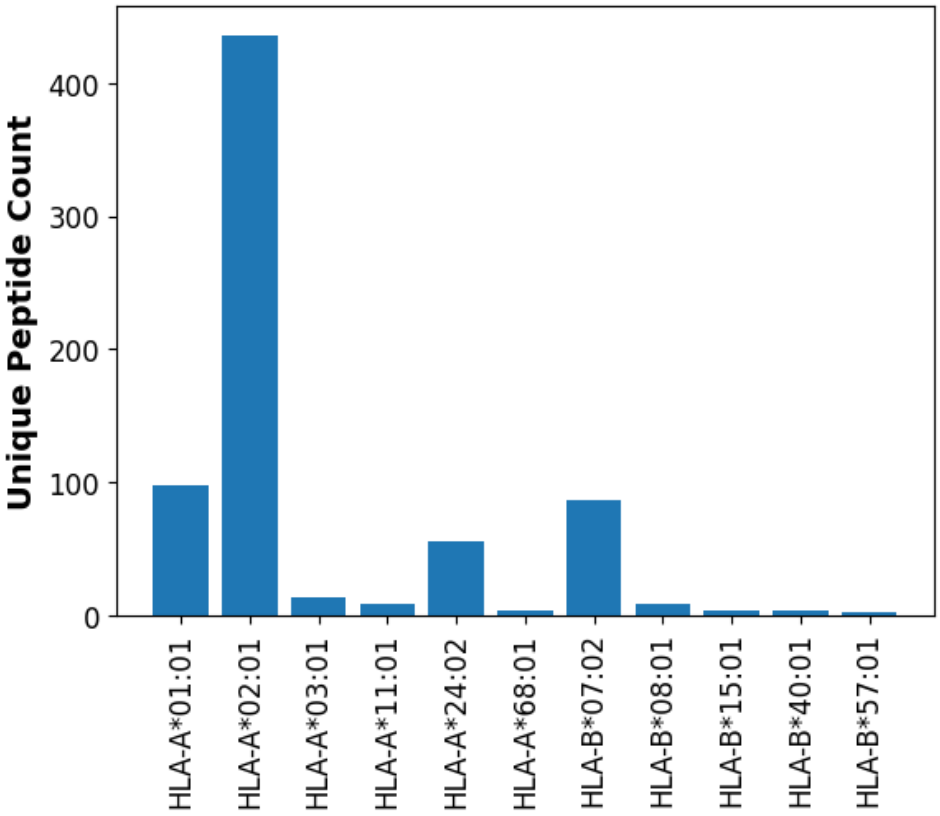

Figure 4.7 The number of unique peptides occurring in TCR-pMHC complexes that contributed to the primary dataset broken down by HLA allele. Note that alleles HLA-A*02:01 and HLA-A*03:01 (columns 2 and 3) have a similar number of records in the dataset but are associated with very different numbers of peptides.

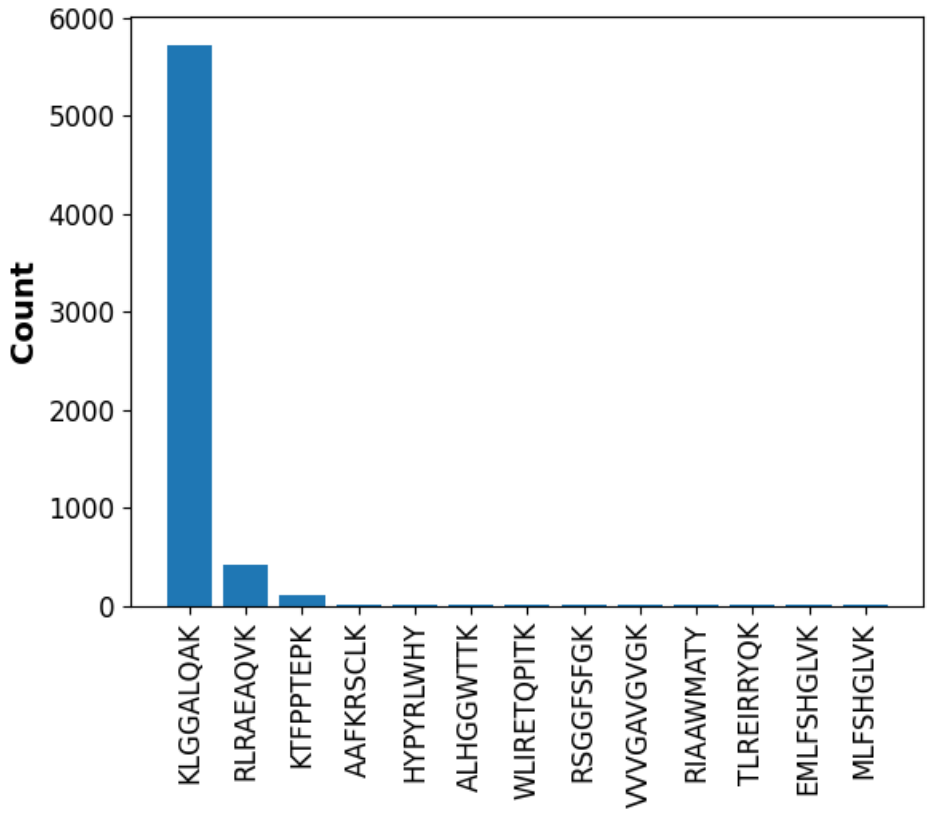

Figure 4.8 TCR-HLA-A*03:01 associations broken down by peptide. Peptides counts are for TCR-pMHC complexes that contributed to the primary dataset having MHC molecules encoded by the HLA-A*03:01 allele. Nearly all TCRs bind to peptide KLGGALQAK.

One interesting feature of the DeepTHAWT results is that the highest AUC scores are for the HLA-A*02:01 allele in both the primary dataset 10-fold cross-validation evaluation (section 4.4.1) and the VDJdb (training) / 10x Genomics (test) evaluation (section 4.4.2). In both cases, TCR-HLA-A*02:01 associations are present in both training and test sets. Yet in the generalised performance (section 4.4.3), the AUC score for HLA-A*02:01 (with TCR-HLA-A*02:01 associations present exclusively in the test set) was by far the lowest – indeed worse than random (see Figure 4.6). At least part of the answer is likely to be that HLA-A*02:01 is special. Firstly, allele group HLA-A*02 occurs at high frequencies in most human populations (see the Allele Frequency Net Database https://allelefrequencies.net/hla.asp), and HLA-A*02:01 is the most common HLA-A*02 allele. Secondly, Blevins and co-workers conclude that human TCRs are “enriched in the capacity to engage” a polymorphic, positively charged ‘hot-spot’ region” that is “almost exclusive” to the α1-helix of HLA-A*02:01-encoded MHC molecules (Blevins et al., 2016). This provides a potential rationale for the discrepancy in performance: only when HLA-A*02:01 is present in the trainings set is a predictor capable of learning about the “almost exclusive” ‘hot-spot’

region. (Whether other HLA alleles are associated with distinctive sequence features is unclear.)

Notwithstanding these points, it has not proved possible to pin down the precise underlying causes for the poor predictive performance with HLA-A*02:01, which remains an unsolved puzzle.

4.5 Correlation analysis

This section investigates the correlation between the HLA-specific AUC scores of DeepTHAWT applied to the primary dataset using 10-fold cross-validation (as described in section 4.4.1) and various relationships between the training and test data evaluated with respect to each of the HLA alleles. In this case, the AUCs are calculated for each HLA allele at each of the ten folds; this was performed using only the six most common HLA alleles in order to ensure robust AUC calculation involving a sufficient number of test set TCR-HLA associations (i.e. at least 50) for every HLA allele.

In summary, the relationships between the HLA-specific test set associations and the training set associations were defined with respect to the following properties:

- The number of HLA-specific training set associations (i.e. TCRs) for the test set HLA allele of interest.
- TCR similarity, which here measures the mean similarity between a) all the test set TCRs associated with the HLA allele of interest and b) all the TCRs in the corresponding training set. As in the correlation analysis for TCR-pMHC binding prediction tools, various properties of the TCR data were explored involving combinations of: one more TCR sequence features (e.g. the CDR3β), with similarities calculated using TCRdist3 (Mayer-Blackwell et al., 2021); and the number of similar training set TCRs to be taken into account.

Note that, with respect to the number of similar training set TCRs that are taken into account, a slightly different strategy was adopted to the one for the correlation analysis of TCR-pMHC binding prediction. Here a training set will typically contain many TCRs that are similar to those in the test set, as both training and test set

contain many TCRs associated with the same HLA allele (this was never the case for the nuTCRacker correlation analysis). Moreover, we are interested in the potential impact of each CDR on prediction accuracy (e.g. does having similar CDR1β sequences in the training set improve prediction accuracy for the HLA alleles in the test set?), and certain CDRs are more variable than others. One consequence is that many training set TCRs may have identical CDR1, 2 and/or 2.5 sequences to ones in the test set, with the proportion of identical sequences varying for different combinations of CDR and HLA. In looking for strong correlations, we should expect that (perhaps very) different proportions of the most similar sequences will prove optimal for different CDRs.

The most highly correlated properties are shown in Figure 4.9. As expected, these differ in important respects from those for the nuTCRacker tool applied to the TCR-pMHC binding prediction problem, for which the key correlations are shown in Figure 3.5. With respect to TCR similarity, only a single TCR feature – the CDR3β – stands out for TCR-pMHC binding prediction, whereas for the prediction of TCR-HLA associations, all CDRs (including the non-classical CDR2.5s) are correlated to moderate degrees. This is broadly in line with expectations, as the CDR3β plays a dominant role in TCR-peptide binding, whereas different combinations of CDRs form contacts with the MHC molecule, depending on the mode of binding (see section 1.4.2).

It is also important that one does not directly interpret differences in the correlation values for different CDRs in terms of their relative biological importance for binding to the MHC molecule. These correlations between the AUC scores and different TCR features are influenced by several factors: by biases in the overall dataset; by the distribution of samples within the folds (given more time and resources, it would have been interesting to investigate how these numbers vary during additional 10-fold cross-validation runs with different random partitioning of the dataset); and by correlations between the degree of similarity observed in different features (e.g. if two sequences have similar CDR2αs, they are likely to have similar CDR2.5αs irrespective of whether either loop is in contact with the MHC molecule).

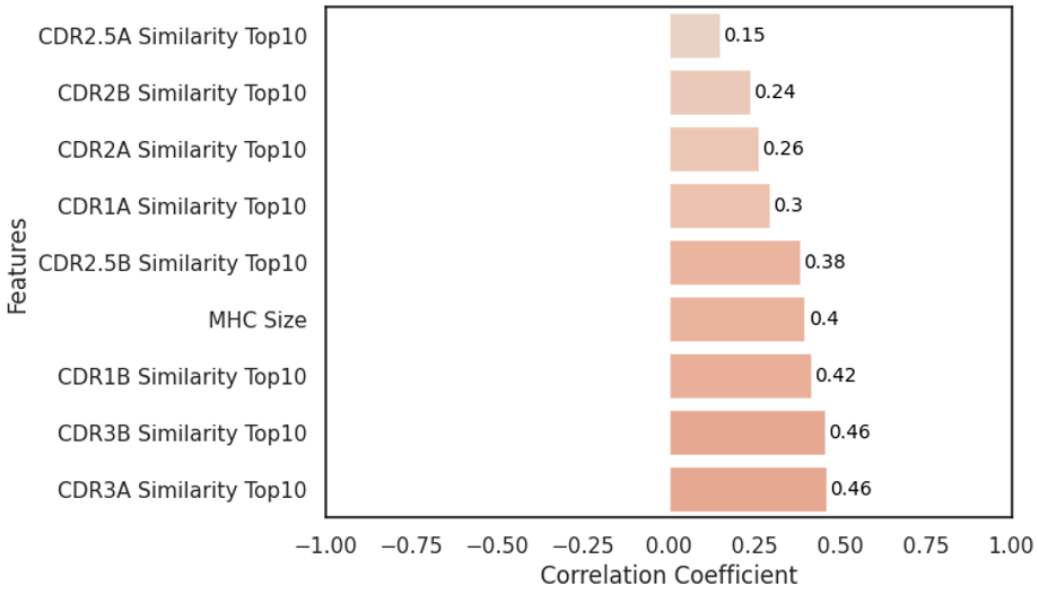

Figure 4.9 Correlogram showing the correlation between AUC and various measures of relationship between training data and HLA-specific test data. MHC Size = the number of training set records having the same HLA allele as the allele of interest in a given fold. For a given type of CDR, Similarity Top x is calculated as follows: firstly, the similarity is measured (using TCRdist3) between a) every sequence of that CDR type belonging to a given HLA/fold combination and b) every sequence of that CDR type in the corresponding dataset; secondly, the similarities are ranked, and the mean of the top x is calculated. There are notable correlations with AUC in all cases.

4.6 Discussion

This chapter has presented a novel transformer-based method, DeepTHAWT, for the prediction of TCR-HLA associations. Using repurposed versions of the dataset developed for benchmarking TCR-pMHC binding prediction tools (Chapter 3), DeepTHAWT achieved AUCs of 0.68 and 0.76 when the same HLAs were present in both training and test sets, but no better than random in the generalised, “unseen” case (with all records for a given HLA present exclusively in the test set).

While this research was being undertaken, two model were published – CLAIRE (Glazer et al., 2022) and DePTH (Liu et al., 2023) – that address the same task using different datasets. Both achieved at least one AUC above 0.8 in the “seen” HLA case, with DePTH additionally reporting AUCs above 0.6 for three “unseen” HLAs.

Taking the AUCs in isolation, DeepTHAWT appears to be the least successful of the three methods. However, as noted in section 4.4.4, CLAIRE was trained on a much larger dataset (its TCR feature requirements are less stringent) and DePTH was

trained using a dataset of public TCRs that may be somewhat easier targets for prediction than TCRs in general. Ultimately, the differences between the datasets used to evaluate these tools pose similar issues to those addressed in Chapter 3 for TCR-pMHC binding prediction tools – specifically, it is unclear how these tools would perform when applied to the same datasets. Even though the task of predicting TCR-HLA associations is in its infancy, with (so far as I am aware) only three methods having been developed to date, it appears we have already reached the point where (preferably independent) benchmarking is needed if we are to understand the relative merits of different predictive tools.

Both DeepTHAWT and CLAIRE performed rather differently when trained and tested on different combinations of dataset. Given the underlying complexities – the different proportions of HLA alleles, the different number of peptides bound to different MHC molecules, the different ways of measuring the binding of TCR-pMHC complexes and setting thresholds, and other potential biases (e.g. towards public TCRs or the selection of strong binders) – it is currently not possible to tease out the impact of different factors. Going forward, it is likely that the choice of dataset needs to be aligned to the intended application. In terms of the application raised at the outset of this chapter, namely the annotation of repertoire datasets, it is perhaps reasonable to conclude that only DePTH has a potential role to play (subject to further testing with different datasets), although one that is limited in terms of the number of HLA alleles it is able to make accurate predictions for and with a bias towards the annotation of public TCRs.

As yet, no tool is capable of making useful predictions for unseen HLA alleles. Whereas it was argued in Chapter 3 that the ability of a TCR-pMHC binding prediction tool to generalise is a key measure of its potential usefulness, this is less true for the task of identifying the likely HLA associated with a given TCR, as (notwithstanding the very large number of HLA alleles discussed in section 4.1) a relatively small number of HLAs afford a large population coverage. As tools emerge that are capable of making accurate predictions for a relatively small number of very common HLA alleles, they are likely to become immediately useful. It is hoped that the research presented in this chapter will make a contribution towards this short-term goal.

4.7 Code availability

DeepTHAWT code is hosted on GitHub: https://github.com/TruptiAG/DeepTHAWT. The repository contains code for tokenising the data and finetuning the model. All the models and datasets are available on our lab's Hugging Face hub here: https://huggingface.co/shepherdgroup.

5 Conclusion

5.1 Main Findings

This thesis has addressed two distinct tasks – the prediction of TCR-pMHC binding and the prediction of TCR-HLA associations – using sequence-based computational prediction methods. The major findings of this research were as follows.
Firstly, in the benchmarking study of TCR-pMHC binding predictors presented in Chapter 3, it was shown that two tools have a moderate ability to generalise to unseen peptides. In particular, it was shown that the new tool nuTCRacker (developed in Prof Adrian Shepherd's group) achieved potentially useful levels of prediction accuracy (with an AUC greater than 0.75) in around a third of cases. This level of performance challenges the widespread assumption that TCR-pMHC binding predictors cannot generalise to unseen peptides. Moreover, although close similarities were found between a few of the unseen antigenic peptides and peptides in the training set, this was not generally the case.

It is, however, important to remain cautious about the wider significance of these results given the inevitable biases of the dataset used to train and test the tools. Key biases are towards high-affinity TCRs bound to peptides from a narrow range of sources (most commonly viruses) and dominated by a small number of HLA alleles.

A second finding was that, by looking at correlations between a) training/test set relationships and b) the corresponding levels of predictive accuracy (in the form of AUC scores), it is possible to suggest (in very broad terms) the minimal requirements for making accurate predictions, namely that the training set contains: a few similar TCRs, with particular reference to their CDR3β sequences; at least one similar peptide; and a large number of samples belonging to the same HLA allele. These are deliberately tentative conclusions, reflecting an awareness of the risk of over-interpretation, particularly given the imbalances and biases inherent in the dataset.

A third contribution was the development of a novel transformer-based method, DeepTHAWT, for the prediction of TCR-HLA associations. DeepTHAWT achieved AUCs of 0.68 and 0.76 when the same HLAs were present in both training and test

sets, but no better than random in the generalised, “unseen” case (with all records for a given HLA present exclusively in the test set). During the course of my PhD, two TCR-HLA association prediction tools were released. Taking all three tools together, it is clear that, even when using the same tool, significantly different levels of performance have been achieved with different datasets. Consequently, the relative merits of the different tools in terms of prediction accuracy is hard to judge.

As with the TCR-pMHC binding predictors nuTCRacker and ERGO II, a correlation analysis was performed. As expected, given our knowledge of the contacts formed within TCR-pMHC complexes (see section 1.4.2), the analysis suggests that a wider set of CDRs make important contributions to the interactions between TCR and MHC than those between TCR and peptide.

5.2 Limitations and Future Work

The research presented in this thesis has several limitations. The first arises from the dataset used for tool benchmarking. Some of the limitations are an inevitable consequence of the current biases found in available data (for example the strong bias towards viral epitopes and a small number of common alleles). Although the most striking imbalances were reduced by downsampling (see sections 3.2.2 and 4.2.1), the datasets used here – and more generally in published research within this field – are inherently biased, which means that conclusions need to be drawn cautiously and with strong caveats concerning their wider applicability.

Different benchmarked tools had different requirements in terms of input features, most notably with respect to TCR sequences. All the samples in the benchmark dataset contained the data required by the most demanding tools, which meant that some samples that were perfectly sufficient for less demanding tools were filtered out. Arguably this put those tools at a disadvantage – they may have performed better with more data, and it can be argued that “fewer features, more data” is a legitimate design choice, given that access to additional training data may outweigh the potential benefits of incorporating additional features. Going forward, it would be useful to explore this trade-off explicitly, although the additional computational burden cannot be ignored.

Bearing in mind the observations in section 2.5 about the bias of certain T cell binding assays towards high-specificity TCRs, it would be useful to explore the impact on predictive performance. For example, to what extent does the inclusion of data derived from MHC tetramers in the training set degrade a tool's ability to make predictions for (less biased) MHC dextramer-derived samples? However, given that there are likely to be other differences between tetramer and dextramer samples (e.g. their respective HLA allele distributions), and given the general lack of data and modest levels of tool performance, it may be infeasible to devise a convincing investigation.

Regarding the analysis of results, the sub-optimality of BLOSUM62 as a means of measuring peptide similarity has already been noted; an alternative that takes into account context-specific constraints (discussed briefly in section 3.5) may be worth investigating. In the current analyses, HLA alleles are treated categorically, with the correlation between AUC, and the number of training MHC samples belonging to exactly the same HLA allele as the test sample MHCs, being one of the strongest correlations with both tasks. To take into account relationships between different but related HLA alleles, an additional "MHC supertype size" could be calculated that represents the number of training set MHCs that belong to the same HLA supertypes (Sidney et al., 2008) as those of the test samples.

Finally, given the involvement of multiple CDRs in the formation of TCR-HLA associations, it may be worth investigating whether there is a correlation between predictive performance and the number of training/test set TCRs sharing the same (or a subset of the same) CDR canonical classes (see section 1.3.2). A method for predicting the CDR canonical classes from a TCR sequence has already been developed (Wong et al., 2019).

5.3 Final Thoughts

At this point, it is worth taking a broader perspective. If the community working in this field wishes to make more than slow, incremental progress, it seems reasonable to conclude that several fundamental issues will need to be addressed. We need much better datasets, and the most efficient way to achieve this would be via high-quality experiments targeted at known gaps in the current pool of available data (for example, with respect to non-viral epitopes and less common HLA alleles). To properly understand which predictive strategies are most effective, there need to be regular, independent benchmarking efforts covering the main tasks of interest and involving multiple datasets. There also needs to be better communication between sequence-based and structure-based researchers, as the most effective future strategies are likely to involve both.

However, any consideration as to what resources the preceding initiatives might require lies firmly outside the scope of this thesis.

Appendix 1

DeepTHAWT Data Encoding

The input data is encoded using Hugging Face Transformers' Tokenizer (2.4.3). It tokenises the input sequences and converts the tokens to their corresponding IDs using a pre-training vocabulary. For example, the input data for CDRs has a vocabulary consisting of the letters for each amino acid residue and its corresponding ID. (e.g. "A": 5,"C": 6,"D": 7,"E": 8…etc). It also has four unique tokens for padding **[PAD]** (padding is needed when the sequences are of different lengths), a separator **[SEP]** between two CDR sequences, a unique token **[UNK]** for unknown amino acids and a special token **[CLS]** inserted at the beginning of the input sequence that stands for classification. The vocabulary also has user-defined and domain-specific custom tokens. The tokens used for this task are: **[cdra1], [cdra2], [cdra25], [cdra3], [cdrb1], [cdrb2], [cdrb25], [cdrb3], [mhc]**. These tokens are used to separate the CDR sequences from each other.

The tokenizer is selected to match the pre-trained model being used for the task, as this ensures that the input sequences are split in the same way as those in the pre-training corpus and uses the same vocabulary. The tokeniser used for this task was the DeBERTa (Decoding-enhanced BERT with disentangled attention) tokenizer. Of the two available options – either a complete Python implementation (Slow version), or a Fast version – the "Fast" version was used. The fast version speeds up the tokenisation process using batch tokenisation and has additional methods that map the original string (in this case a sequence of amino acids) with the token index. This aids the reconstruction of the original "text" (in this case a set of CDR sequences) for visualisation and debugging purposes.

After parsing the input, the tokenizer returns a dictionary where **input_ids** are indices corresponding to each token in the sequence (Figure A1).

```
tokenizer("[cdra1]SVFSS[cdra2]VVTGGEV[cdrb1]SGHRS[cdrb2]YFSETQ
]")

{'input_ids': [2, 28, 20, 22, 9, 20, 20, 29, 22, 22, 21, 10,
10, 8, 22, 25, 20, 10, 11, 19, 20, 26, 24, 9, 20, 8, 21, 18,
3],
```

Figure A1 Tokenisation. The first line shows the input sequences consisting of a CDR1a, a CDR2a , a CDR1b and a CDR2b, interspersed with the domain-specific custom tokens [cdra1], [cdra2], [cdrb1] and [cdrb2]. (The CDR3 and MHC sequences are not shown in the figure.) The input_ids are the indices for corresponding tokens. The first token, ID 2, represents the start of sequence token [CLS], and token ID 3 represents the end of sequence token [SEP]. All other tokens are for the amino acids in the CDR sequences and the domain-specific custom tokens, for example token ID 28 is for the token [cdra1] and token ID 20 is for amino acid ‘S’.

References

Abanades, B., Wong, W. K., Boyles, F., Georges, G., Bujotzek, A., & Deane, C. M. (2023). ImmuneBuilder: Deep-Learning models for predicting the structures of immune proteins. *Communications Biology*, *6*(1), 575. https://doi.org/10.1038/s42003-023-04927-7

Abi Habib, J., Lesenfants, J., Vigneron, N., & Van den Eynde, B. J. (2022). Functional differences between proteasome subtypes. *Cells (Basel, Switzerland)*, *11*(3), 421. https://doi.org/10.3390/cells11030421

Archbold, J. K., Macdonald, W. A., Gras, S., Ely, L. K., Miles, J. J., Bell, M. J., Brennan, R. M., Beddoe, T., Wilce, M. C. J., Clements, C. S., Purcell, A. W., McCluskey, J., Burrows, S. R., & Rossjohn, J. (2009). Natural micropolymorphism in human leukocyte antigens provides a basis for genetic control of antigen recognition. *The Journal of Experimental Medicine*, *206*(1), 209–219. https://doi.org/10.1084/jem.20082136

Armstrong, K. M., Piepenbrink, K. H., & Baker, B. M. (2008). Conformational changes and flexibility in T-cell receptor recognition of peptide-MHC complexes. *The Biochemical Journal*, *415*(2), 183–196. https://doi.org/10.1042/BJ20080850

Aronson, A., Hochner, T., Cohen, T., & Schneidman-Duhovny, D. (2022). Structure modeling and specificity of peptide-MHC class I interactions using geometric deep learning. In: *Bioinformatics* (biorxiv;2022.12.15.520566v1). bioRxiv. https://www.biorxiv.org/content/10.1101/2022.12.15.520566v1.full.pdf

Arstila, T. P., Casrouge, A., Baron, V., Even, J., Kanellopoulos, J., & Kourilsky, P. (1999). A direct estimate of the human alphabeta T cell receptor diversity. *Science (New York, N.Y.)*, *286*(5441), 958–961. https://doi.org/10.1126/science.286.5441.958

Atchley, W. R., Zhao, J., Fernandes, A. D., & Drüke, T. (2005). Solving the protein sequence metric problem. *Proceedings of the National Academy of Sciences of the*

United States of America, *102*(18), 6395–6400. https://doi.org/10.1073/pnas.0408677102

Baker, B. M., Scott, D. R., Blevins, S. J., & Hawse, W. F. (2012). Structural and dynamic control of T-cell receptor specificity, cross-reactivity, and binding mechanism. *Immunological Reviews*, *250*(1), 10–31. https://doi.org/10.1111/j.1600-065x.2012.01165.x

Barbosa, C. R. R., Barton, J., Shepherd, A. J., & Mishto, M. (2021). Mechanistic diversity in MHC class I antigen recognition. *Biochemical Journal*, *478*(24), 4187–4202. https://doi.org/10.1042/BCJ20200910

Barker, D. J., Maccari, G., Georgiou, X., Cooper, M. A., Flicek, P., Robinson, J., & Marsh, S. G. E. (2023). The IPD-IMGT/HLA database. *Nucleic Acids Research*, *51*(D1), D1053–D1060. https://doi.org/10.1093/nar/gkac1011

Blevins, S. J., Pierce, B. G., Singh, N. K., Riley, T. P., Wang, Y., Spear, T. T., Nishimura, M. I., Weng, Z., & Baker, B. M. (2016). How structural adaptability exists alongside HLA-A2 bias in the human αβ TCR repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(9), E1276-85. https://doi.org/10.1073/pnas.1522069113

Borbulevych, O. Y., Piepenbrink, K. H., Gloor, B. E., Scott, D. R., Sommese, R. F., Cole, D. K., Sewell, A. K., & Baker, B. M. (2009). T cell receptor cross-reactivity directed by antigen-dependent tuning of peptide-MHC molecular flexibility. *Immunity*, *31*(6), 885–896. https://doi.org/10.1016/j.immuni.2009.11.003

Borg, N. A., Ely, L. K., Beddoe, T., Macdonald, W. A., Reid, H. H., Clements, C. S., Purcell, A. W., Kjer-Nielsen, L., Miles, J. J., Burrows, S. R., McCluskey, J., & Rossjohn, J. (2005). The CDR3 regions of an immunodominant T cell receptor dictate the "energetic landscape" of peptide-MHC recognition. *Nature Immunology*, *6*(2), 171–180. https://doi.org/10.1038/ni1155

Bradley, P. (2023). Structure-based prediction of T cell receptor:peptide-MHC interactions. *ELife*, *12*. https://doi.org/10.7554/eLife.82813

Burrows, S. R., Chen, Z., Archbold, J. K., Tynan, F. E., Beddoe, T., Kjer-Nielsen, L., Miles, J. J., Khanna, R., Moss, D. J., Liu, Y. C., Gras, S., Kostenko, L., Brennan, R. M., Clements, C. S., Brooks, A. G., Purcell, A. W., McCluskey, J., & Rossjohn, J. (2010). Hard wiring of T cell receptor specificity for the major histocompatibility complex is underpinned by TCR adaptability. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(23), 10608–10613. https://doi.org/10.1073/pnas.1004926107

Cai, M., Bang, S., Zhang, P., & Lee, H. (2022). ATM-TCR: TCR-Epitope Binding Affinity Prediction Using a Multi-Head Self-Attention Model. *Frontiers in Immunology*, *13*, 893247. https://doi.org/10.3389/fimmu.2022.893247

Castorina, L. V., Grazioli, F., Machart, P., Mösch, A., & Errica, F. (2023). Assessing the Generalization Capabilities of TCR Binding Predictors via Peptide Distance Analysis. In: *Immunology* (biorxiv;2023.07.29.551100v2). bioRxiv. https://www.biorxiv.org/content/10.1101/2023.07.29.551100v2.full.pdf

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase representations using RNN encoder-decoder for statistical machine translation. In: *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1406.1078

Chronister, W. D., Crinklaw, A., Mahajan, S., Vita, R., Koşaloğlu-Yalçın, Z., Yan, Z., Greenbaum, J. A., Jessen, L. E., Nielsen, M., Christley, S., Cowell, L. G., Sette, A., & Peters, B. (2021). TCRMatch: Predicting T-cell receptor specificity based on sequence similarity to previously characterized receptors. *Frontiers in Immunology*, *12*, 640725. https://doi.org/10.3389/fimmu.2021.640725

Colbert, J. D., Cruz, F. M., & Rock, K. L. (2020). Cross-presentation of exogenous antigens on MHC I molecules. *Current Opinion in Immunology*, *64*, 1–8. https://doi.org/10.1016/j.coi.2019.12.005

Corrie, B. D., Marthandan, N., Zimonja, B., Jaglale, J., Zhou, Y., Barr, E., Knoetze, N., Breden, F. M. W., Christley, S., Scott, J. K., Cowell, L. G., & Breden, F. (2018). iReceptor: A platform for querying and analyzing antibody/B-cell and T-cell receptor repertoire data across federated repositories. *Immunological Reviews*, *284*(1), 24–41. https://doi.org/10.1111/imr.12666

Dash, P., Fiore-Gartland, A. J., Hertz, T., Wang, G. C., Sharma, S., Souquette, A., Crawford, J. C., Clemens, E. B., Nguyen, T. H. O., Kedzierska, K., La Gruta, N. L., Bradley, P., & Thomas, P. G. (2017). Quantifiable predictive features define epitope-specific T cell receptor repertoires. *Nature*, *547*(7661), 89–93. https://doi.org/10.1038/nature22383

De Boer, R. J., & Perelson, A. S. (2013). Quantifying T lymphocyte turnover. *Journal of Theoretical Biology*, *327*, 45–87. https://doi.org/10.1016/j.jtbi.2012.12.025

Deng, L., Ly, C., Abdollahi, S., Zhao, Y., Prinz, I., & Bonn, S. (2023). Performance comparison of TCR-pMHC prediction tools reveals a strong data dependency. *Frontiers in Immunology*, *14*, 1128326. https://doi.org/10.3389/fimmu.2023.1128326

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1810.04805

DeWitt, W. S., 3rd, Smith, A., Schoch, G., Hansen, J. A., Matsen, F. A., 4th, & Bradley, P. (2018). Human T cell receptor occurrence patterns encode immune history, genetic background, and receptor specificity. *ELife*, *7*. https://doi.org/10.7554/eLife.38358

Dolton, G., Lissina, A., Skowera, A., Ladell, K., Tungatt, K., Jones, E., Kronenberg-Versteeg, D., Akpovwa, H., Pentier, J. M., Holland, C. J., Godkin, A. J., Cole, D. K., Neller, M. A., Miles, J. J., Price, D. A., Peakman, M., & Sewell, A. K. (2014). Comparison of peptide-major histocompatibility complex tetramers and dextramers for the identification of antigen-specific T cells. *Clinical and Experimental Immunology*, *177*(1), 47–63. https://doi.org/10.1111/cei.12339

Dunbar, J., & Deane, C. M. (2016). ANARCI: antigen receptor numbering and receptor classification. *Bioinformatics* , *32*(2), 298–300. https://doi.org/10.1093/bioinformatics/btv552

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2021). ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Learning. In: *bioRxiv* (p. 2020.07.12.199554). https://doi.org/10.1101/2020.07.12.199554

Emerson, R. O., DeWitt, W. S., Vignali, M., Gravley, J., Hu, J. K., Osborne, E. J., Desmarais, C., Klinger, M., Carlson, C. S., Hansen, J. A., Rieder, M., & Robins, H. S. (2017). Immunosequencing identifies signatures of cytomegalovirus exposure history and HLA-mediated effects on the T cell repertoire. *Nature Genetics*, *49*(5), 659–665. https://doi.org/10.1038/ng.3822

Evans, R., O'Neill, M., Pritzel, A., Antropova, N., Senior, A., Green, T., Žídek, A., Bates, R., Blackwell, S., Yim, J., Ronneberger, O., Bodenstein, S., Zielinski, M., Bridgland, A., Potapenko, A., Cowie, A., Tunyasuvunakool, K., Jain, R., Clancy, E., … Hassabis, D. (2021). Protein complex prediction with AlphaFold-Multimer. In: *Bioinformatics* (biorxiv;2021.10.04.463034v2). bioRxiv. https://www.biorxiv.org/content/10.1101/2021.10.04.463034v2.full.pdf

Fang, Y., Liu, X., & Liu, H. (2022). Attention-aware contrastive learning for predicting T cell receptor-antigen binding specificity. *Briefings in Bioinformatics*, *23*(6). https://doi.org/10.1093/bib/bbac378

Fischer, D. S., Wu, Y., Schubert, B., & Theis, F. J. (2020). Predicting antigen specificity of single T cells based on TCR CDR3 regions. *Molecular Systems Biology*, *16*(8), e9416. https://doi.org/10.15252/msb.20199416

Gao, Y., Gao, Y., Dong, K., Wu, S., & Liu, Q. (2023). Reply to: The pitfalls of negative data bias for the T-cell epitope specificity challenge. *Nature Machine Intelligence*, *5*(10), 1063–1065. https://doi.org/10.1038/s42256-023-00725-2

Gielis, S., Moris, P., Bittremieux, W., De Neuter, N., Ogunjimi, B., Laukens, K., & Meysman, P. (2019). Detection of Enriched T Cell Epitope Specificity in Full T Cell Receptor

Sequence Repertoires. *Frontiers in Immunology*, *10*, 2820. https://doi.org/10.3389/fimmu.2019.02820

Gil, A., Kamga, L., Chirravuri-Venkata, R., Aslan, N., Clark, F., Ghersi, D., Luzuriaga, K., & Selin, L. K. (2020). Epstein-Barr virus Epitope-major histocompatibility complex interaction combined with convergent recombination drives selection of diverse T cell receptor α and β repertoires. *MBio*, *11*(2). https://doi.org/10.1128/mBio.00250-20

Gilbert, S. C. (2013). Advances in the development of universal influenza vaccines. *Influenza and Other Respiratory Viruses*, *7*(5), 750–758. https://doi.org/10.1111/irv.12013

Glanville, J., Huang, H., Nau, A., Hatton, O., Wagar, L. E., Rubelt, F., Ji, X., Han, A., Krams, S. M., Pettus, C., Haas, N., Arlehamn, C. S. L., Sette, A., Boyd, S. D., Scriba, T. J., Martinez, O. M., & Davis, M. M. (2017). Identifying specificity groups in the T cell receptor repertoire. *Nature*, *547*(7661), 94–98. https://doi.org/10.1038/nature22976

Glazer, N., Akerman, O., & Louzoun, Y. (2022). Naive and memory T cells TCR-HLA-binding prediction. *Oxford Open Immunology*, *3*(1), iqac001. https://doi.org/10.1093/oxfimm/iqac001

Gonzalez-Galarza, F. F., McCabe, A., Santos, E. J. M. D., Jones, J., Takeshita, L., Ortega-Rivera, N. D., Cid-Pavon, G. M. D., Ramsbottom, K., Ghattaoraya, G., Alfirevic, A., Middleton, D., & Jones, A. R. (2020). Allele frequency net database (AFND) 2020 update: gold-standard data classification, open access genotype data and new query tools. *Nucleic Acids Research*, *48*(D1), D783–D788. https://doi.org/10.1093/nar/gkz1029

Gras, S., Chadderton, J., Del Campo, C. M., Farenc, C., Wiede, F., Josephs, T. M., Sng, X. Y. X., Mirams, M., Watson, K. A., Tiganis, T., Quinn, K. M., Rossjohn, J., & La Gruta, N. L. (2016). Reversed T cell receptor docking on a major histocompatibility class I complex limits involvement in the immune response. *Immunity*, *45*(4), 749–760. https://doi.org/10.1016/j.immuni.2016.09.007

Grazioli, F., Mösch, A., Machart, P., Li, K., Alqassem, I., O'Donnell, T. J., & Min, M. R. (2022). On TCR binding predictors failing to generalize to unseen peptides. *Frontiers in Immunology*, *13*, 1014256. https://doi.org/10.3389/fimmu.2022.1014256

Harndahl, M., Rasmussen, M., Roder, G., Dalgaard Pedersen, I., Sørensen, M., Nielsen, M., & Buus, S. (2012). Peptide-MHC class I stability is a better predictor than peptide affinity of CTL immunogenicity: Antigen processing. *European Journal of Immunology*, *42*(6), 1405–1416. https://doi.org/10.1002/eji.201141774

He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. In: *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/2006.03654

Heather, J. M., Ismail, M., Oakes, T., & Chain, B. (2018). High-throughput sequencing of the T-cell receptor repertoire: pitfalls and opportunities. *Briefings in Bioinformatics*, *19*(4), 554–565. https://doi.org/10.1093/bib/bbw138

Ho, T. K. (1995). Random decision forests. *Proceedings of 3rd International Conference on Document Analysis and Recognition*, *1*, 278–282 vol.1. https://doi.org/10.1109/ICDAR.1995.598994

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735

Holtmeier, W., & Kabelitz, D. (2005). gammadelta T cells link innate and adaptive immune responses. *Chemical Immunology and Allergy*, *86*, 151–183. https://doi.org/10.1159/000086659

Huang, H., Wang, C., Rubelt, F., Scriba, T. J., & Davis, M. M. (2020). Analyzing the Mycobacterium tuberculosis immune response by T-cell receptor clustering with GLIPH2 and genome-wide antigen screening. *Nature Biotechnology*, *38*(10), 1194–1202. https://doi.org/10.1038/s41587-020-0505-4

Hudson, D., Fernandes, R. A., Basham, M., Ogg, G., & Koohy, H. (2023). Can we predict T cell specificity with digital biology and machine learning? *Nature Reviews. Immunology*, 1–11. https://doi.org/10.1038/s41577-023-00835-3

Jiang, Y., Huo, M., & Cheng Li, S. (2023). TEINet: a deep learning framework for prediction of TCR-epitope binding specificity. *Briefings in Bioinformatics*, *24*(2). https://doi.org/10.1093/bib/bbad086

Jokinen, E., Huuhtanen, J., Mustjoki, S., Heinonen, M., & Lähdesmäki, H. (2021). Predicting recognition between T cell receptors and epitopes with TCRGP. *PLoS Computational Biology*, *17*(3), e1008814. https://doi.org/10.1371/journal.pcbi.1008814

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaas, Q., Ruiz, M., & Lefranc, M.-P. (2004). IMGT/3Dstructure-DB and IMGT/StructuralQuery, a database and a tool for immunoglobulin, T cell receptor and MHC structural data. *Nucleic Acids Research*, *32*(Database issue), D208-10. https://doi.org/10.1093/nar/gkh042

Khan, A. R., Reinders, M. J. T., & Khatri, I. (2023). Determining epitope specificity of T-cell receptors with transformers. *Bioinformatics* , *39*(11). https://doi.org/10.1093/bioinformatics/btad632

Khan, J. M., & Ranganathan, S. (2011). Understanding TR binding to pMHC complexes: how does a TR scan many pMHC complexes yet preferentially bind to one. *PloS One*, *6*(2), e17194. https://doi.org/10.1371/journal.pone.0017194

Kwee, B. P. Y., Messemaker, M., Marcus, E., Oliveira, G., Scheper, W., Wu, C., Teuwen, J., & Schumacher, T. (2023). STAPLER: Efficient learning of TCR-peptide specificity prediction from full-length TCR-peptide data. In: *bioRxiv* (p. 2023.04.25.538237). https://doi.org/10.1101/2023.04.25.538237

La Gruta, N. L., Gras, S., Daley, S. R., Thomas, P. G., & Rossjohn, J. (2018). Understanding the drivers of MHC restriction of T cell receptors. *Nature Reviews. Immunology*, *18*(7), 467–478. https://doi.org/10.1038/s41577-018-0007-5

Lee, W., & Suresh, M. (2022). Vaccine adjuvants to engage the cross-presentation pathway. *Frontiers in Immunology*, *13*, 940047. https://doi.org/10.3389/fimmu.2022.940047

Leem, J., de Oliveira, S. H. P., Krawczyk, K., & Deane, C. M. (2018). STCRDab: the structural T-cell receptor database. *Nucleic Acids Research*, *46*(D1), D406–D412. https://doi.org/10.1093/nar/gkx971

Liu, S., Bradley, P., & Sun, W. (2023). Neural network models for sequence-based TCR and HLA association prediction. *PLoS Computational Biology*, *19*(11), e1011664. https://doi.org/10.1371/journal.pcbi.1011664

Lu, T., Zhang, Z., Zhu, J., Wang, Y., Jiang, P., Xiao, X., Bernatchez, C., Heymach, J. V., Gibbons, D. L., Wang, J., Xu, L., Reuben, A., & Wang, T. (2021). Deep learning-based prediction of the T cell receptor–antigen binding specificity. *Nature Machine Intelligence*, *3*(10), 864–875. https://doi.org/10.1038/s42256-021-00383-2

Mackie, P. L. (2003). The classification of viruses infecting the respiratory tract. *Paediatric Respiratory Reviews*, *4*(2), 84–90. https://doi.org/10.1016/s1526-0542(03)00031-9

Madeira, F., Madhusoodanan, N., Lee, J., Eusebi, A., Niewielska, A., Tivey, A. R. N., Lopez, R., & Butcher, S. (2024). The EMBL-EBI Job Dispatcher sequence analysis tools framework in 2024. *Nucleic Acids Research*, *52*(W1), W521–W525. https://doi.org/10.1093/nar/gkae241

Manso, T., Folch, G., Giudicelli, V., Jabado-Michaloud, J., Kushwaha, A., Nguefack Ngoune, V., Georga, M., Papadaki, A., Debbagh, C., Pégorier, P., Bertignac, M., Hadi-Saljoqi, S., Chentli, I., Cherouali, K., Aouinti, S., El Hamwi, A., Albani, A., Elazami Elhassani, M., Viart, B., … Kossida, S. (2022). IMGT® databases, related tools and web resources through three main axes of research and development. *Nucleic Acids Research*, *50*(D1), D1262–D1272. https://doi.org/10.1093/nar/gkab1136

Mason, D. (1998). A very high level of crossreactivity is an essential feature of the T-cell receptor. *Immunology Today*, *19*(9), 395–404. https://doi.org/10.1016/s0167-5699(98)01299-7

Mayer-Blackwell, K., Schattgen, S., Cohen-Lavi, L., Crawford, J. C., Souquette, A., Gaevert, J. A., Hertz, T., Thomas, P. G., Bradley, P., & Fiore-Gartland, A. (2021). TCR meta-clonotypes for biomarker discovery with tcrdist3 enabled identification of public, HLA-restricted clusters of SARS-CoV-2 TCRs. *ELife*, *10*. https://doi.org/10.7554/eLife.68605

McDaniel, M. M., Meibers, H. E., & Pasare, C. (2021). Innate control of adaptive immunity and adaptive instruction of innate immunity: bi-directional flow of information. *Current Opinion in Immunology*, *73*, 25–33. https://doi.org/10.1016/j.coi.2021.07.013

McMaster, B., Thorpe, C. J., Rossjohn, J., Deane, C. M., & Koohy, H. (2024). Quantifying conformational changes in the TCR:pMHC-I binding interface. *Frontiers in Immunology*, *15*, 1491656. https://doi.org/10.3389/fimmu.2024.1491656

McMaster, B., Thorpe, C., Ogg, G., Deane, C. M., & Koohy, H. (2024). Can AlphaFold's breakthrough in protein structure help decode the fundamental principles of adaptive cellular immunity? *Nature Methods*, *21*(5), 766–776. https://doi.org/10.1038/s41592-024-02240-7

Meyer, S., Blaas, I., Bollineni, R. C., Delic-Sarac, M., Tran, T. T., Knetter, C., Dai, K.-Z., Madssen, T. S., Vaage, J. T., Gustavsen, A., Yang, W., Nissen-Meyer, L. S. H., Douvlataniotis, K., Laos, M., Nielsen, M. M., Thiede, B., Søraas, A., Lund-Johansen, F., Rustad, E. H., & Olweus, J. (2023). Prevalent and immunodominant CD8 T cell epitopes are conserved in SARS-CoV-2 variants. *Cell Reports*, *42*(1), 111995. https://doi.org/10.1016/j.celrep.2023.111995

Meysman, P., Barton, J., Bravi, B., Cohen-Lavi, L., Karnaukhov, V., Lilleskov, E., Montemurro, A., Nielsen, M., Mora, T., Pereira, P., Postovskaya, A., Martínez, M. R., Fernandez-de-Cossio-Diaz, J., Vujkovic, A., Walczak, A. M., Weber, A., Yin, R., Eugster, A., & Sharma, V. (2023). Benchmarking solutions to the T-cell receptor

epitope prediction problem: IMMREP22 workshop report. *Immunoinformatics (Amsterdam, Netherlands)*, *9*(100024), 100024. https://doi.org/10.1016/j.immuno.2023.100024

Montemurro, A., Jessen, L. E., & Nielsen, M. (2022). NetTCR-2.1: Lessons and guidance on how to develop models for TCR specificity predictions. *Frontiers in Immunology*, *13*, 1055151. https://doi.org/10.3389/fimmu.2022.1055151

Montemurro, A., Schuster, V., Povlsen, H. R., Bentzen, A. K., Jurtz, V., Chronister, W. D., Crinklaw, A., Hadrup, S. R., Winther, O., Peters, B., Jessen, L. E., & Nielsen, M. (2021). NetTCR-2.0 enables accurate prediction of TCR-peptide binding by using paired TCRα and β sequence data. *Communications Biology*, *4*(1), 1060. https://doi.org/10.1038/s42003-021-02610-3

Moris, P., De Pauw, J., Postovskaya, A., Gielis, S., De Neuter, N., Bittremieux, W., Ogunjimi, B., Laukens, K., & Meysman, P. (2021). Current challenges for unseen-epitope TCR interaction prediction and a new perspective derived from image classification. *Briefings in Bioinformatics*, *22*(4). https://doi.org/10.1093/bib/bbaa318

Murphy, K., & Weaver, C. (2017). *Janeway's immunobiology 9th edition*. Garland Science. https://inmunologos.wordpress.com/wp-content/uploads/2020/08/janeways-immunobiology-9th-ed_booksmedicos.org_.pdf

Nielsen, M., Lundegaard, C., Blicher, T., Lamberth, K., Harndahl, M., Justesen, S., Røder, G., Peters, B., Sette, A., Lund, O., & Buus, S. (2007). NetMHCpan, a method for quantitative predictions of peptide binding to any HLA-A and -B locus protein of known sequence. *PloS One*, *2*(8), e796. https://doi.org/10.1371/journal.pone.0000796

Omer, A., Peres, A., Rodriguez, O. L., Watson, C. T., Lees, W., Polak, P., Collins, A. M., & Yaari, G. (2022). T cell receptor beta germline variability is revealed by inference from repertoire data. *Genome Medicine*, *14*(1), 2. https://doi.org/10.1186/s13073-021-01008-4

Ortega, M. R., Pogorelyy, M., Minervina, A., Thomas, P., Walczak, A., & Mora, T. (2024). Learning predictive signatures of HLA type from T-cell repertoires. In: *bioRxiv* (p. 2024.01.25.577228). https://doi.org/10.1101/2024.01.25.577228

Paul, W. E. (2012). The immune system – complexity exemplified. *Mathematical Modelling of Natural Phenomena*, *7*(5), 4–6. https://doi.org/10.1051/mmnp/20127502

Peacock, T., & Chain, B. (2021). Information-driven docking for TCR-pMHC complex prediction. *Frontiers in Immunology*, *12*, 686127. https://doi.org/10.3389/fimmu.2021.686127

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research: JMLR*, *12*(85), 2825–2830. http://jmlr.org/papers/v12/pedregosa11a.html

Penn, D. J., Damjanovich, K., & Potts, W. K. (2002). MHC heterozygosity confers a selective advantage against multiple-strain infections. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(17), 11260–11264. https://doi.org/10.1073/pnas.162006499

Perez, M. A. S., Cuendet, M. A., Röhrig, U. F., Michielin, O., & Zoete, V. (2022). Structural prediction of peptide-MHC binding modes. *Methods in Molecular Biology (Clifton, N.J.)*, *2405*, 245–282. https://doi.org/10.1007/978-1-0716-1855-4_13

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. https://doi.org/10.1002/jcc.20084

Pham, M.-D. N., Nguyen, T.-N., Tran, L. S., Nguyen, Q.-T. B., Nguyen, T.-P. H., Pham, T. M. Q., Nguyen, H.-N., Giang, H., Phan, M.-D., & Nguyen, V. (2023). epiTCR: a highly sensitive predictor for TCR-peptide binding. *Bioinformatics* , *39*(5). https://doi.org/10.1093/bioinformatics/btad284

Pierce, B. G., & Weng, Z. (2013). A flexible docking approach for prediction of T cell receptor-peptide-MHC complexes. *Protein Science: A Publication of the Protein Society*, *22*(1), 35–46. https://doi.org/10.1002/pro.2181

Povlsen, H. R., Bentzen, A. K., Kadivar, M., Jessen, L. E., Hadrup, S. R., & Nielsen, M. (2023). Improved T cell receptor antigen pairing through data-driven filtering of sequencing information from single cells. *ELife*, *12*. https://doi.org/10.7554/eLife.81810

Qi, Q., Liu, Y., Cheng, Y., Glanville, J., Zhang, D., Lee, J.-Y., Olshen, R. A., Weyand, C. M., Boyd, S. D., & Goronzy, J. J. (2014). Diversity and clonal selection in the human T-cell repertoire. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(36), 13139–13144. https://doi.org/10.1073/pnas.1409155111

Ramsuran, V., Kulkarni, S., O'huigin, C., Yuki, Y., Augusto, D. G., Gao, X., & Carrington, M. (2015). Epigenetic regulation of differential HLA-A allelic expression levels. *Human Molecular Genetics*, *24*(15), 4268–4275. https://doi.org/10.1093/hmg/ddv158

Rapin, N., Hoof, I., Lund, O., & Nielsen, M. (2008). MHC motif viewer. *Immunogenetics*, *60*(12), 759–765. https://doi.org/10.1007/s00251-008-0330-2

Reynisson, B., Alvarez, B., Paul, S., Peters, B., & Nielsen, M. (2020). NetMHCpan-4.1 and NetMHCIIpan-4.0: improved predictions of MHC antigen presentation by concurrent motif deconvolution and integration of MS MHC eluted ligand data. *Nucleic Acids Research*, *48*(W1), W449–W454. https://doi.org/10.1093/nar/gkaa379

Rudolph, M. G., Stanfield, R. L., & Wilson, I. A. (2006). How TCRs bind MHCs, peptides, and coreceptors. *Annual Review of Immunology*, *24*, 419–466. https://doi.org/10.1146/annurev.immunol.23.021704.115658

Saper, M., Bjorkman, P., & Wiley, D. (1991). Refined structure of the human histocompatibility antigen HLA-A2 at 2.6 A resolution. *Journal of Molecular Biology*, *219*(2), 277–319. https://doi.org/10.1016/0022-2836(91)90567-P

Schodin, B. A., Tsomides, T. J., & Kranz, D. M. (1996). Correlation between the number of T cell receptors required for T cell activation and TCR-ligand affinity. *Immunity*, *5*(2), 137–146. https://doi.org/10.1016/s1074-7613(00)80490-2

Schuster, M., Paliwal, K. K., & Member. (1997). Bidirectional Recurrent Neural Networks. *IEEE Transactions on Signal Processing: A Publication of the IEEE Signal Processing Society*, *45*(11).

Sewell, A. K. (2012). Why must T cells be cross-reactive? *Nature Reviews. Immunology*, *12*(9), 669–677. https://doi.org/10.1038/nri3279

Shcherbinin, D. S., Belousov, V. A., & Shugay, M. (2020). Comprehensive analysis of structural and sequencing data reveals almost unconstrained chain pairing in TCRαβ complex. *PLoS Computational Biology*, *16*(3), e1007714. https://doi.org/10.1371/journal.pcbi.1007714

Shugay, M., Bagaev, D. V., Zvyagin, I. V., Vroomans, R. M., Crawford, J. C., Dolton, G., Komech, E. A., Sycheva, A. L., Koneva, A. E., Egorov, E. S., Eliseev, A. V., Van Dyk, E., Dash, P., Attaf, M., Rius, C., Ladell, K., McLaren, J. E., Matthews, K. K., Clemens, E. B., … Chudakov, D. M. (2018). VDJdb: a curated database of T-cell receptor sequences with known antigen specificity. *Nucleic Acids Research*, *46*(D1), D419–D427. https://doi.org/10.1093/nar/gkx760

Sidhom, J.-W., Larman, H. B., Pardoll, D. M., & Baras, A. S. (2021). DeepTCR is a deep learning framework for revealing sequence concepts within T-cell repertoires. *Nature Communications*, *12*(1), 1605. https://doi.org/10.1038/s41467-021-21879-w

Sidney, J., Peters, B., Frahm, N., Brander, C., & Sette, A. (2008). HLA class I supertypes: a revised and updated classification. *BMC Immunology*, *9*, 1. https://doi.org/10.1186/1471-2172-9-1

Slade, R. W., & McCallum, H. I. (1992). Overdominant vs. frequency-dependent selection at MHC loci. *Genetics*, *132*(3), 861–864. https://doi.org/10.1093/genetics/132.3.861

Smith, A. R., Alonso, J. A., Ayres, C. M., Singh, N. K., Hellman, L. M., & Baker, B. M. (2021). Structurally silent peptide anchor modifications allosterically modulate T cell recognition in a receptor-dependent manner. *Proceedings of the National Academy*

of Sciences of the United States of America, *118*(4), e2018125118. https://doi.org/10.1073/pnas.2018125118

Springer, I., Tickotsky, N., & Louzoun, Y. (2021). Contribution of T Cell Receptor Alpha and Beta CDR3, MHC Typing, V and J Genes to Peptide Binding Prediction. *Frontiers in Immunology*, *12*, 664514. https://doi.org/10.3389/fimmu.2021.664514

Spurgin, L. G., & Richardson, D. S. (2010). How pathogens drive genetic diversity: MHC, mechanisms and misunderstandings. *Proceedings. Biological Sciences*, *277*(1684), 979–988. https://doi.org/10.1098/rspb.2009.2084

Steinegger, M., Mirdita, M., & Söding, J. (2019). Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, *16*(7), 603–606. https://doi.org/10.1038/s41592-019-0437-4

Sun, Y., Iglesias, E., Samri, A., Kamkamidze, G., Decoville, T., Carcelain, G., & Autran, B. (2003). A systematic comparison of methods to measure HIV-1 specific CD8 T cells. *Journal of Immunological Methods*, *272*, 23–34.

Sun, Yimo, Li, F., Sonnemann, H., Jackson, K. R., Talukder, A. H., Katailiha, A. S., & Lizee, G. (2021). Evolution of CD8+ T cell receptor (TCR) engineered therapies for the treatment of cancer. *Cells (Basel, Switzerland)*, *10*(9), 2379. https://doi.org/10.3390/cells10092379

Szeto, C., Lobos, C. A., Nguyen, A. T., & Gras, S. (2020). TCR Recognition of Peptide-MHC-I: Rule Makers and Breakers. *International Journal of Molecular Sciences*, *22*(1). https://doi.org/10.3390/ijms22010068

Tadros, D. M., Eggenschwiler, S., Racle, J., & Gfeller, D. (2022). The MHC Motif Atlas: a database of MHC binding specificities and ligands. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkac965

Tickotsky, N., Sagiv, T., Prilusky, J., Shifrut, E., & Friedman, N. (2017). McPAS-TCR: a manually curated catalogue of pathology-associated T cell receptor sequences. *Bioinformatics* , *33*(18), 2924–2929. https://doi.org/10.1093/bioinformatics/btx286

Tong, Y., Wang, J., Zheng, T., Zhang, X., Xiao, X., Zhu, X., Lai, X., & Liu, X. (2020). SETE: Sequence-based Ensemble learning approach for TCR Epitope binding prediction.

Computational Biology and Chemistry, *87*, 107281. https://doi.org/10.1016/j.compbiolchem.2020.107281

Trolle, T., McMurtrey, C. P., Sidney, J., Bardet, W., Osborn, S. C., Kaever, T., Sette, A., Hildebrand, W. H., Nielsen, M., & Peters, B. (2016). The length distribution of class I-restricted T cell epitopes is determined by both peptide supply and MHC allele-specific binding preference. *The Journal of Immunology*, *196*(4), 1480–1487. https://doi.org/10.4049/jimmunol.1501721

Tynan, F. E., Burrows, S. R., Buckle, A. M., Clements, C. S., Borg, N. A., Miles, J. J., Beddoe, T., Whisstock, J. C., Wilce, M. C., Silins, S. L., Burrows, J. M., Kjer-Nielsen, L., Kostenko, L., Purcell, A. W., McCluskey, J., & Rossjohn, J. (2005). T cell receptor recognition of a "super-bulged" major histocompatibility complex class I-bound peptide. *Nature Immunology*, *6*(11), 1114–1122. https://doi.org/10.1038/ni1257

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In: *arXiv [cs.CL]*. arXiv. http://arxiv.org/abs/1706.03762

Venturi, V., Kedzierska, K., Price, D. A., Doherty, P. C., Douek, D. C., Turner, S. J., & Davenport, M. P. (2006). Sharing of T cell receptors in antigen-specific responses is driven by convergent recombination. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(49), 18691–18696. https://doi.org/10.1073/pnas.0608907103

Venturi, V., Quigley, M. F., Greenaway, H. Y., Ng, P. C., Ende, Z. S., McIntosh, T., Asher, T. E., Almeida, J. R., Levy, S., Price, D. A., Davenport, M. P., & Douek, D. C. (2011). A mechanism for TCR sharing between T cell subsets and individuals revealed by pyrosequencing. *The Journal of Immunology*, *186*(7), 4285–4294. https://doi.org/10.4049/jimmunol.1003898

Vermijlen, D., Gatti, D., Kouzeli, A., Rus, T., & Eberl, M. (2018). γδ T cell responses: How many ligands will it take till we know? *Seminars in Cell & Developmental Biology*, *84*, 75–86. https://doi.org/10.1016/j.semcdb.2017.10.009

Vita, R., Mahajan, S., Overton, J. A., Dhanda, S. K., Martini, S., Cantrell, J. R., Wheeler, D. K., Sette, A., & Peters, B. (2019). *The Immune Epitope Database (IEDB)*. https://www.iedb.org

Wculek, S. K., Cueto, F. J., Mujal, A. M., Melero, I., Krummel, M. F., & Sancho, D. (2020). Dendritic cells in cancer immunology and immunotherapy. *Nature Reviews. Immunology*, *20*(1), 7–24. https://doi.org/10.1038/s41577-019-0210-z

Weber, A., Born, J., & Rodriguez Martínez, M. (2021). TITAN: T-cell receptor specificity prediction with bimodal attention networks. *Bioinformatics* , *37*(Suppl_1), i237–i244. https://doi.org/10.1093/bioinformatics/btab294

Weber, A., Pélissier, A., & Rodríguez Martínez, M. (2024). T-cell receptor binding prediction: A machine learning revolution. *Immunoinformatics (Amsterdam, Netherlands)*, *100040*, 100040. https://doi.org/10.1016/j.immuno.2024.100040

Winternitz, J., Abbate, J. L., Huchard, E., Havlíček, J., & Garamszegi, L. Z. (2017). Patterns of MHC-dependent mate selection in humans and nonhuman primates: a meta-analysis. *Molecular Ecology*, *26*(2), 668–688. https://doi.org/10.1111/mec.13920

Wong, W. K., Leem, J., & Deane, C. M. (2019). Comparative analysis of the CDR loops of antigen receptors. *Frontiers in Immunology*, *10*, 2454. https://doi.org/10.3389/fimmu.2019.02454

Wooldridge, L., Ekeruche-Makinde, J., van den Berg, H. A., Skowera, A., Miles, J. J., Tan, M. P., Dolton, G., Clement, M., Llewellyn-Lacey, S., Price, D. A., Peakman, M., & Sewell, A. K. (2012). A single autoimmune T cell receptor recognizes more than a million different peptides. *The Journal of Biological Chemistry*, *287*(2), 1168–1177. https://doi.org/10.1074/jbc.M111.289488

Wu, D., Yin, R., Chen, G., Ribeiro-Filho, H. V., Cheung, M., Robbins, P. F., Mariuzza, R. A., & Pierce, B. G. (2024). Structural characterization and AlphaFold modeling of human T cell receptor recognition of NRAS cancer neoantigens. In: *Immunology* (biorxiv;2024.05.21.595215v1). bioRxiv. https://www.biorxiv.org/content/10.1101/2024.05.21.595215v1.full.pdf

Wu, J., Qi, M., Zhang, F., & Zheng, Y. (2023). TPBTE: A model based on convolutional Transformer for predicting the binding of TCR to epitope. *Molecular Immunology*, *157*, 30–41. https://doi.org/10.1016/j.molimm.2023.03.010

Wu, K., Yost, K. E., Daniel, B., Belk, J. A., Xia, Y., Egawa, T., Satpathy, A., Chang, H. Y., & Zou, J. (2021). TCR-BERT: learning the grammar of T-cell receptors for flexible antigen-xbinding analyses. In: *bioRxiv* (p. 2021.11.18.469186). https://doi.org/10.1101/2021.11.18.469186

Yewdell, J. W., Reits, E., & Neefjes, J. (2003). Making sense of mass destruction: quantitating MHC class I antigen presentation. *Nature Reviews. Immunology*, *3*(12), 952–961. https://doi.org/10.1038/nri1250

Yin, R., Ribeiro-Filho, H. V., Lin, V., Gowthaman, R., Cheung, M., & Pierce, B. G. (2023). TCRmodel2: high-resolution modeling of T cell receptor recognition using deep learning. *Nucleic Acids Research*, *51*(W1), W569–W576. https://doi.org/10.1093/nar/gkad356

Yoo, S., Jeong, M., Seomun, S., Kim, K., & Han, Y. (2024). Interpretable Prediction of SARS-CoV-2 Epitope-specific TCR Recognition Using a Pre-Trained Protein Language Model. *IEEE/ACM Transactions on Computational Biology and Bioinformatics / IEEE, ACM*, *PP*. https://doi.org/10.1109/TCBB.2024.3368046

Yu, K., Shi, J., Lu, D., & Yang, Q. (2019). Comparative analysis of CDR3 regions in paired human αβ CD8 T cells. *FEBS Open Bio*, *9*(8), 1450–1459. https://doi.org/10.1002/2211-5463.12690

Yu, X. G., Lichterfeld, M., Chetty, S., Williams, K. L., Mui, S. K., Miura, T., Frahm, N., Feeney, M. E., Tang, Y., Pereyra, F., LaBute, M. X., Pfafferott, K., Leslie, A., Crawford, H., Allgaier, R., Hildebrand, W., Kaslow, R., Brander, C., Allen, T. M., … Walker, B. D. (2007). Mutually exclusive T-cell receptor induction and differential susceptibility to human immunodeficiency virus type 1 mutational escape associated with a two-amino-acid difference between HLA class I subtypes. *Journal of Virology*, *81*(4), 1619–1631. https://doi.org/10.1128/jvi.01580-06

Zhang, S., Bakshi, R. K., Suneetha, P. V., Fytili, P., Antunes, D. A., Vieira, G. F., Jacobs, R., Klade, C. S., Manns, M. P., Kraft, A. R. M., Wedemeyer, H., Schlaphoff, V., & Cornberg, M. (2015). Frequency, private specificity, and cross-reactivity of preexisting hepatitis C virus (HCV)-specific CD8+ T cells in HCV-seronegative individuals: Implications for vaccine responses. *Journal of Virology*, *89*(16), 8304–8317. https://doi.org/10.1128/JVI.00539-15

Zhang, Wei, Wang, L., Liu, K., Wei, X., Yang, K., Du, W., Wang, S., Guo, N., Ma, C., Luo, L., Wu, J., Lin, L., Yang, F., Gao, F., Wang, X., Li, T., Zhang, R., Saksena, N. K., Yang, H., … Liu, X. (2020). PIRD: Pan Immune Repertoire Database. *Bioinformatics* , *36*(3), 897–903. https://doi.org/10.1093/bioinformatics/btz614

Zhang, Wen, Hawkins, P. G., He, J., Gupta, N. T., Liu, J., Choonoo, G., Jeong, S. W., Chen, C. R., Dhanik, A., Dillon, M., Deering, R., Macdonald, L. E., Thurston, G., & Atwal, G. S. (2021). A framework for highly multiplexed dextramer mapping and prediction of T cell receptor sequences to antigen specificity. *Science Advances*, *7*(20). https://doi.org/10.1126/sciadv.abf5835

Zhao, Y., He, B., Xu, F., Li, C., Xu, Z., Su, X., He, H., Huang, Y., Rossjohn, J., Song, J., & Yao, J. (2023). DeepAIR: A deep learning framework for effective integration of sequence and 3D structure to enable adaptive immune receptor analysis. *Science Advances*, *9*(32), eabo5128. https://doi.org/10.1126/sciadv.abo5128