In silico Characterisation of Antigen Receptor Binding Site Structures

Wing Ki Wong

New College
University of Oxford

A thesis submitted for the degree of
Doctor of Philosophy

Michaelmas 2020

Abstract

The adaptive immune system defends the host against the invasion of foreign molecules known as the "antigens". These antigens are recognised by two main types of antigen receptors: T-cell receptors (TCRs) and antibodies. While both proteins share a globally similar β-sandwich architecture and are encoded by similar genetic mechanisms, TCRs are polyspecific and have medium affinity to their peptide antigens, while antibodies are highly specific and have high affinity to their targets. Their different behaviours are thought to be at least partially dictated by their binding site features. In this thesis, we aim to analyse their binding site structures and develop tools that can leverage the greater breadth of binding site diversity unveiled by repertoire sequencing.

In both types of proteins, the majority of the binding site is constituted by the complementarity-determining region (CDR) loops. In this thesis we first describe the development of a rapid sequence-based canonical form prediction tool (SCALOP) for antibody CDRs and for TCR CDRs. Based on this initial structural annotation, we then explored the structural differences between antibody and TCR CDRs and found that TCR CDRs tend to adopt multiple conformations, more often than their antibody counterparts. To capture the potential ensemble of binding site conformations, we built a TCR modelling tool, TCRBuilder.

Moving on from the structure of CDR alone, we then developed Ab-Ligity, a structure-based method that identifies sequence-dissimilar antibodies against the same epitope. This method incorporates the predicted antibody structure and the physicochemical properties of the binding site. Finally, as Ab-Ligity is dependent upon prediction of the paratope, we evaluated the leading paratope prediction method, Parapred. We attempt to highlight possible features that could portray paratopes with interpretable statistical models. The thesis concludes with the potential future directions of the work on binding site analysis for antibodies and TCRs.

In silico Characterisation of Antigen Receptor Binding Site Structures

Wing Ki Wong

New College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2020

For my parents, who have made a lot of sacrifices to support my dream.
For my brother, who taught me patience.

Acknowledgements

I have been incredibly fortunate throughout my DPhil study, with the support from many individuals. I would like to start by first thanking my supervisor, Professor Charlotte M. Deane, who has trusted me and given me the opportunity to pursue a doctorate with her. She has offered invaluable advice and guidance on my work over the past four years, generously sharing her wisdom and teaching me by example how to be an independent researcher. I am privileged to have her as my supervisor. I would also like to thank my large team of industrial supervisors: Guy Georges and Alexander Bujotzek from Roche Diagnostics, Alan P. Lewis from GlaxoSmithKline, Jiye Shi and Sebastian Kelm from UCB, and Bojana Popovic and Bruck Taddese from AstraZeneca (formerly MedImmune). It was helpful to improve my work after implementing it in an industrial setting and hearing useful feedback. I am grateful for the studentship funding from EPSRC and MRC, as well as additional support from the four pharmaceutical companies.

Having companions throughout the journey was important to me, especially when I have been motivated by the high productivity of my fellow antibody OPIGlets, Aleksandr Kovaltsuk and Matthew Raybould. It has been amazing to meet a diverse team of immunoinformatics researchers in OPIG. My special thanks go to Jinwoo Leem and Konrad Krawczyk for their encouragement at the start of my DPhil; and to Claire Marks, Constantin Schneider, Eve Richardson and Sarah Robinson for stimulating a holistic thinking about research in this field.

Outside of research, I am most indebted to my friends in SysBio, SABS and the wider Oxford community. Casey Adam, Anne Nierobisch and James Wilsenach have been great mates who are always there for me, and who are compassionate towards my situations. Jacky Lei has consistently brought delights to my days since our first academic term. My New College social buddies, Khairulanwar Zaini and Rachel Qiu, have seen all my ups-and-downs and sent their support from afar. My OxFEST 2018 – 2019 committee is a group of superwomen who has brought joy in the midst of pursuing a virtuous goal of equal opportunity across genders in STEM. I am humbled to be the leader of such an outstanding group, to complete my experience at Oxford.

To my girls from undergraduate and high school: Yan Ting Lau-Woodcock, Angel Tseung, Flora Lee, Christina Lee and Susannah Lau. You reminded me that I had come a long way, and I am sincerely grateful for your company through my different life stages. To my partner Leung Sing Chan, you made me a happier and better person.

Last but not least, a heartfelt thank you to my family – my parents for their hard work that brought me to where I am, and my brother for keeping up with me. This DPhil is not only for me, but also for all of you.

Abstract

The adaptive immune system defends the host against the invasion of foreign molecules known as the "antigens". These antigens are recognised by two main types of antigen receptors: T-cell receptors (TCRs) and antibodies. While both proteins share a globally similar β-sandwich architecture and are encoded by similar genetic mechanisms, TCRs are polyspecific and have medium affinity to their peptide antigens, while antibodies are highly specific and have high affinity to their targets. Their different behaviours are thought to be at least partially dictated by their binding site features. In this thesis, we aim to analyse their binding site structures and develop tools that can leverage the greater breadth of binding site diversity unveiled by repertoire sequencing.

In both types of proteins, the majority of the binding site is constituted by the complementarity-determining region (CDR) loops. In this thesis we first describe the development of a rapid sequence-based canonical form prediction tool (SCALOP) for antibody CDRs and for TCR CDRs. Based on this initial structural annotation, we then explored the structural differences between antibody and TCR CDRs and found that TCR CDRs tend to adopt multiple conformations, more often than their antibody counterparts. To capture the potential ensemble of binding site conformations, we built a TCR modelling tool, TCRBuilder.

Moving on from the structure of CDR alone, we then developed Ab-Ligity, a structure-based method that identifies sequence-dissimilar antibodies against the same epitope. This method incorporates the predicted antibody structure and the physicochemical properties of the binding site. Finally, as Ab-Ligity is dependent upon prediction of the paratope, we evaluated the leading paratope prediction method, Parapred. We attempt to highlight possible features that could portray paratopes with interpretable statistical models. The thesis concludes with the potential future directions of the work on binding site analysis for antibodies and TCRs.

Declaration

I declare that no parts of this thesis or its research herein have been reproduced or accepted for another award or degree or diploma at any other university or learning institution. This thesis contains no other person's work except where stated in the text. I have used both first-person singular ("I") and plural ("we") forms in this thesis interchangeably to indicate work for which I have played a major role in terms of preparation, execution and analysis.

Wing Ki Wong
Michaelmas 2020

Contents

List of Figures

List of Tables

List of Publications and Preprints

Publications

- W. K. Wong, C. Marks, J. Leem, A. P. Lewis, J. Shi, and C. M. Deane. TCRBuilder: multi-state T-cell receptor structure prediction. *Bioinformatics*, 36(11):3580–3581, 2020.
- A. Kovaltsuk, M. I. J. Raybould, W. K. Wong, C. Marks, S. Kelm, J. Snowden, J. Trück,and C. M. Deane. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLoS Computational Biology*, 16(2):e1007636, 2020.
- W. K. Wong, J. Leem, and C. M. Deane. Comparative analysis of the CDR loops of antigen receptors. *Frontiers in Immunology*, 10:2454, 2019.
- J.-P. Ebejer, P. W. Finn, W. K. Wong, C. M. Deane, and G. M. Morris. Ligity: A Non-Superpositional, Knowledge-Based Approach to Virtual Screening. *Journal of Chemical Information and Modeling*, 59(6):2600–2616, 2019.
- M. I. J. Raybould[*], W. K. Wong[*], and C. M. Deane. Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Molecular Systems Design & Engineering*, 4(4):679–688, 2019. ([*]Joint first authors)
- W. K. Wong, G. Georges, F. Ros, S. Kelm, A. P. Lewis, B. Taddese, J. Leem, and C. M. Deane. SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics*, 35(10):1774–1776, 2019.
- K. Krawczyk, S. Kelm, A. Kovaltsuk, J. D. Galson, D. Kelly, J. Trück, C. Regep, J. Leem, W. K. Wong, J. Nowak, J. Snowden, M. Wright, L. Starkie, A. Scott-Tucker,J. Shi, and C. M. Deane. Structurally mapping antibody repertoires. *Frontiers in Immunology*, 9:1698, 2018.

Preprints

- W. K. Wong, S. A. Robinson, A. Bujotzek, G. Georges, A. P. Lewis, J. Shi, J. Snowden, B. Taddese, and C. M. Deane. Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. *BioRxiv*, 2020.
- Y. Wang, A. Tsitsiklis, W. Gao, H. H. Chu, Y. Zhang, W. Li, W. K. Wong, C. M. Deane, D. Neau, J. E. Slansky, P. G. Thomas, E. A. Robey and S. Dai. Novel Vβ specific germline contacts shape an elite controller T cell response. *BioRxiv*, 2020.

1 Introduction

Contents

In this chapter, the section on paratope prediction was adapted from a joint first author review published in *Molecular Systems Design and Engineering* (Raybould et al., 2019b). In this publication, I wrote the majority of the sections on paratope and epitope predictions, and on complex modelling that are used in the text below.

1.1 Introduction

In humans, the adaptive immune system defends the host against a wide range of foreign molecules, or "antigens". These antigens are recognised by two types of receptors: T-cell receptors (TCRs) and antibodies (Janeway et al., 2001). Despite the similar genetic mechanisms and architecture in these two antigen receptors (Dunbar et al., 2014a; Sillitoe et al., 2015), they carry out distinct biological functions and

exhibit different binding properties. This distinction is thought to be driven by differences in their antigen-binding sites. Understanding their binding site properties is critical to inform how to benefit from the desirable binding profiles of both types of proteins to design molecular machines (Nguyen et al., 2017).

The aim of this thesis is to study the binding site structures of antibodies and TCRs. This introductory chapter covers an overview of antibodies and TCRs, and the existing methods to model their binding sites and interactions with the antigen. We start with the structural and sequence characteristics of antibodies and TCRs, and elaborate on the similarity between their genetic mechanisms and architecture. Next we briefly mention the rise of high-throughput immunoglobulin sequencing, how the volume of sequencing data adds breadth to our understanding of antibody space, and how it is possible to leverage structural knowledge to enrich sequence analysis and study antigen receptor binding. We then describe the common experimental assays used to characterise antibody-antigen binding in early-stage antibody discovery, along with the complementary computational models that have been employed to predict binding sites and interactions. This chapter finishes with an overview of the thesis chapters.

1.2 Structural diversity in antigen receptors

The two main types of antigen receptors, TCRs and antibodies, have different biological functions but similar architecture. TCRs typically recognise peptide antigens presented via the major histocompatibility complex (MHC; Rudolph et al., 2006), while antibodies can bind almost any types of antigen, including proteins, peptides and haptens (Sela-Culang et al., 2013). TCRs are known to be polyspecific (*i.e.* one TCR can bind to different peptides) and they tend to have lower affinity to their antigen (Rudolph et al., 2006). On the other hand, antibodies can bind to their target antigen with high specificity and affinity (Alberts et al., 2002; Sela-Culang et al., 2013), making them an ideal class of molecules for diagnostic and

therapeutic applications (Chames et al., 2009). Despite their different roles in the immune response, these proteins share a β-sandwich fold (Figure 1.1; Dunbar et al., 2014a; Sillitoe et al., 2015). Below we describe each of these receptors in more detail.

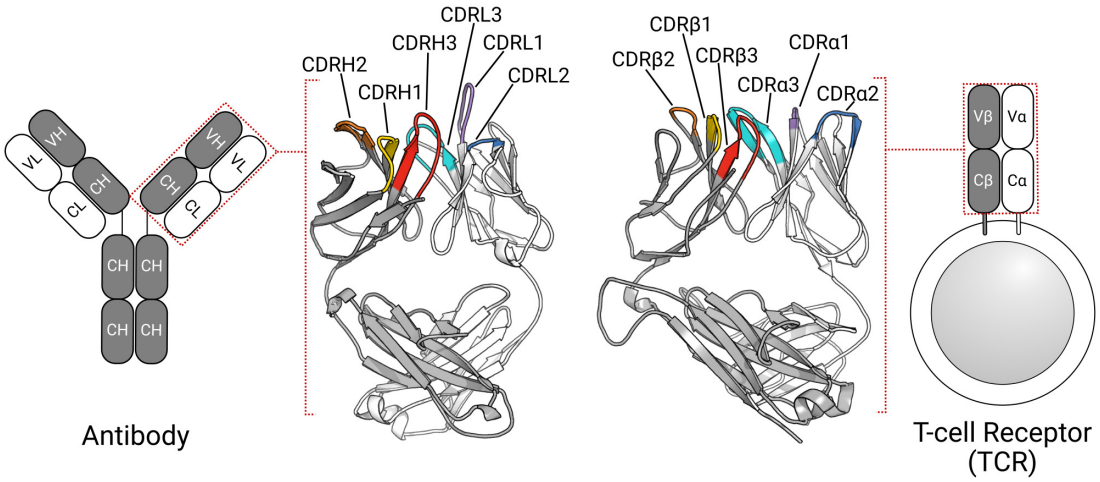


Figure 1.1: Illustrations of T-cell receptor (TCR) and antibody structures. TCRs and antibodies share a globally similar structure. Both proteins are heterodimers, characterised by a set of six CDRs that form the majority of the binding site. Comparable chains and CDRs share colouring schemes; for example, the TCRβ and antibody heavy chain are coloured grey, while CDRα1 and CDRL1 are coloured purple. Figure reproduced from (Wong et al., 2019b).

1.2.1 Antibodies

Antibodies are produced by B-cells in the immune system (Alberts et al., 2002). It was thought that each B-cell could produce only a single species of antibody, but recently this notion has been challenged as Shi et al. (2019) found that multiple heavy chains may be encoded by a single B-cell. The antibody presented on the B-cell surface is known as the B-cell receptor (BCR). When a naive or memory B-cell is activated by an antigen and proliferates, it differentiates and becomes an effector cell that secretes soluble antibodies (Alberts et al., 2002). In the remainder of this thesis, we use "antibody" and "B-cell receptor" interchangeably.

A standard human antibody has a Y-shaped configuration composed of two pairs of heavy and light chains (Figure 1.2A and B). The antigen-binding fragments (Fab) are located on the two "arms". Within the Fab region, the variable domains of the heavy and light chains (collectively known as the Fv region) have the greatest sequence diversity between different antibodies. In particular, sequence variation is concentrated in the six hypervariable loops, known as the complementarity determining regions (CDRs). The three CDRs in the heavy chain are called H1, H2 and H3, whereas those in the light chain are L1, L2 and L3. The CDRs form the majority of the antibody binding site and largely determine the specificity and affinity of the antibody to its cognate antigen (Kabat and Wu, 1991).

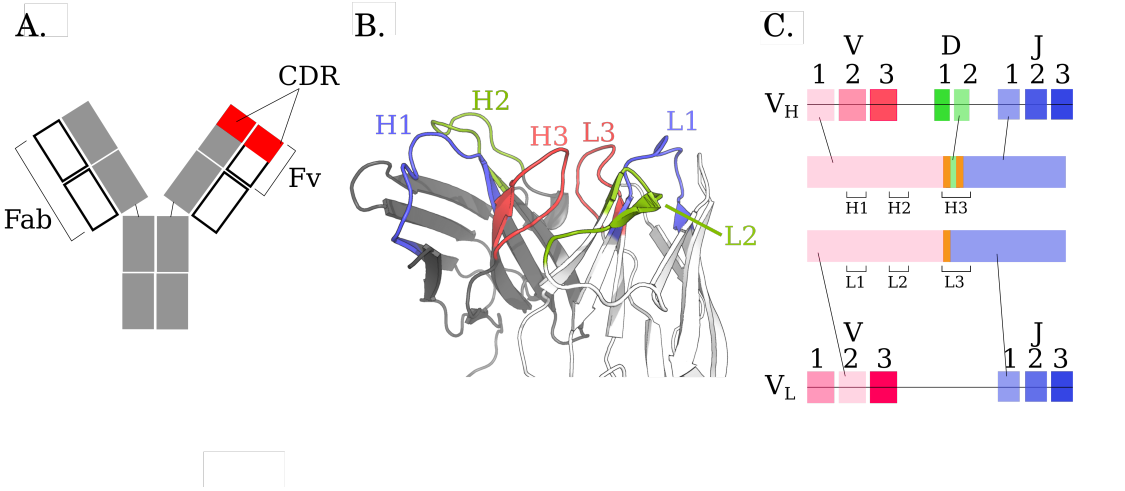


Figure 1.2: A standard human antibody structure and V(D)J recombination. **A.** A schematic diagram of a standard human antibody. Fab and Fv regions are labelled. The heavy chains are in grey, light chains in white and CDRs are in red. **B.** A cartoon representation of the Fv region of a paired antibody. Blue, green and red loops represent the first, second and third hypervariable loops (*i.e.* the CDRs) on each of heavy and light chains. The framework of the heavy chain is coloured in grey and that of the light chain is coloured in white. **C.** A schematic diagram of the V(D)J recombination. V, D and J gene variants are represented by different shades of pink, green and purple respectively. The orange regions are the junctions where the segments join: addition and deletion occur at these junctions, introducing "junctional diversity". The CDR-encoding regions are labelled with the respective CDR types; somatic hypermutation (SHM) concentrates in these regions (Avnir et al., 2014; Francica et al., 2015).

1.2.1.1 Sequence diversity in antibodies

The sequence diversity of the Fv regions arises from V(D)J recombination and somatic hypermutation (SHM; Alberts et al., 2002). The random joining of the V, D and J genes for heavy chain (V and J for light chain) and the random addition and deletion of nucleotides at the junctions ("junctional diversity") diversify the genes encoding the third hypervariable loops of each chain – H3 and L3 (Figure 1.2C; Janeway et al., 2001). Spontaneous mutations and insertions-deletions ("indel") in SHM can introduce further sequence diversification in the other CDRs (Avnir et al., 2014; Francica et al., 2015), leading to a wide range of binding site structures (Tiller and Tessier, 2015).

1.2.1.2 The characteristics of antibody-antigen binding sites

In order to understand the underlying mechanisms of antibody-antigen recognition, the binding sites of antibodies and antigens have been analysed to identify key features for their binding. The antigen-binding residues on the antibody form the "paratope" (Figure 1.6), the bulk of which is from the CDRs (Kunik et al., 2012b). The CDRs have been extensively studied and a plethora of structural prediction tools for these hypervariable loops has been developed (see Section 1.2.1.3). Since the CDRs only broadly define the paratopes, there has been significant interest in pinpointing the exact residues involved in antigen recognition. The mainstream sequence-based and structure-based paratope prediction tools will be discussed in Section 1.4.2.1.

An epitope of an antigen is defined as a subset of its surface to which an antibody binds (Figure 1.6). Epitopes can be divided into two categories: linear and conformational epitopes. Linear epitopes are continuous sequences, typically found in peptide antigens. This class of epitopes can be recognised by its primary structure (Pai et al., 2011). Conformational epitopes are groups of non-sequential

residues that are proximal in three-dimensional space (Najar et al., 2017).

The gold standard method for determining paratopes and epitopes is the inspection of antibody-antigen co-crystal complexes (Martin and Thornton, 1996; Nguyen et al., 2017). However, crystal structures are often hard to obtain and not scalable (Chiu et al., 2019). Other *in vitro* assays have served as surrogates to characterise the binding landscape of a large set of antibodies at early stage antibody screening. For instance, competition assays can cluster antibodies by their epitopic specificity; while small-angle scattering, mass-spectrometry-based techniques and mutagenesis followed by kinetics analysis can identify binding regions at varying levels of resolutions. These methods will be discussed further in Section 1.4.1. Complementing these wet-lab approaches, computational techniques have been used to help identify antibodies for more refined *in vitro* experimental validations. Sequence features, binding site structures and energetics have inspired the development of various descriptors (see Section 1.4.2).

1.2.1.3 CDR structures

One of the most common ways to analyse antibody binding sites is to focus on the CDRs. The six CDR loops make up the majority of the binding site. Five of the six CDRs (except CDRH3) have a limited set of conformations known as the "canonical forms" (see an example in Figure 1.3). This model was first proposed for antibodies in 1987 (Chothia and Lesk, 1987) using only five antibody structures and clusters were defined manually. Martin and Thornton (1996) clustered the CDRs in the torsional space by hierarchical models and derived sequence templates automatically for each cluster. North et al. (2011) also clustered in torsional space but with an affinity propagation algorithm. All of these methods first stratified the CDR structures by their sequence lengths. However, Nowak et al. (2016) found that CDRs with different lengths could fall into the same structural cluster. The authors developed a length-independent clustering scheme with the dynamic time warping

algorithm, to extract CDR canonical forms. The definition of canonical classes has been revisited multiple times as more structures become available (*e.g.* Chothia and Lesk, 1987; Martin and Thornton, 1996; Kuroda et al., 2009; North et al., 2011; Nowak et al., 2016). Canonical forms have been used for *in silico* antibody design (Pantazes and Maranas, 2010; Lapidoth et al., 2015). Sequence features, such as the presence of specific amino acids within or near the CDR loop, may be used to predict the canonical forms (*e.g.* Martin and Thornton, 1996; North et al., 2011; Nowak et al., 2016), and has been used to predict the structures of CDR sequences from immunoglobulin gene sequencing (Ig-seq) datasets (Nowak et al., 2016; Kovaltsuk et al., 2020).

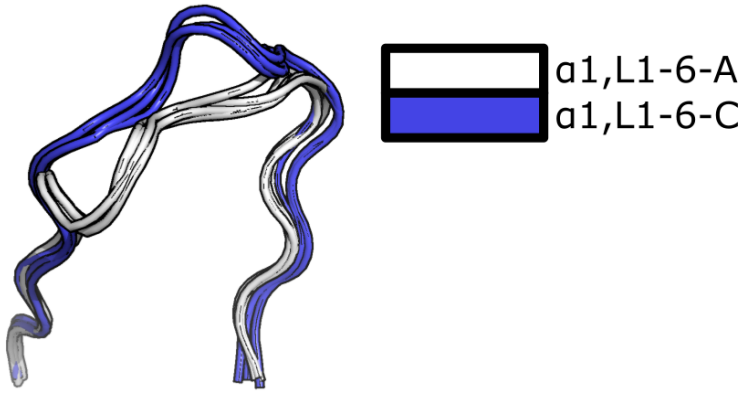


Figure 1.3: An illustration of two CDRL1/α1 canonical forms. CDRL1/*alpha*1 structures from two canonical forms (α1,L1-6-A in white and α1,L1-6-C in blue) were superimposed, aligning five residues before and five after the CDR. See Chapter 3 for the nomenclature of the canonical forms.

In the case of CDRH3, analysis of its anchors has found that the majority of CDRH3 loops adopt a "kinked" base (Weitzner et al., 2015; Finn et al., 2016). The conformation of the central part of the loop is highly variable, and its computational structural prediction remains a challenging area for all current types of methods in database-search, *ab initio* and *de novo* modelling (*e.g.* Choi and Deane, 2010; Leem et al., 2016; Weitzner et al., 2017; Marks et al., 2017). Aside from the crystallographic conformation, the dynamics of the CDR3 loops on heavy and light

chains have been studied with molecular dynamic (MD) simulations that attempt to reproduce the potential loop movements in solutions (Fernández-Quintero et al., 2019a,b). These studies covered specific cases to find out the potential role of loop motions in antibody binding and affinity maturation. These MD analyses have reported differing results on the correlation between structural rigidity and binding affinity, depending on factors such as the simulation timescale and whether the datasets were from crystal structures or models (Jeliazkov et al., 2018; Fernández-Quintero et al., 2020a).

1.2.1.4 Physicochemical properties of the binding surfaces

Human proteins including antibodies and TCRs are formed of 20 *L*-amino acids (Janeway et al., 2001), that share the same backbone composition (C_a, C, N and O), but different side chains. These side chains govern the physicochemical features of the paratopes and epitopes (Janeway et al., 2001). We will first outline the physicochemical properties of each of these amino acids, then describe existing views on the binding site features.

The 20 amino acids have distinct side chains. Evolutionary analysis suggested that some residue substitutions are more favourable and may functionally replace another, and these statistical potentials have been summarised in matrices such as BLOSUM and PAM (Wilbur, 1985; Henikoff and Henikoff, 1992). Similarity between their physicochemical properties such as hydrophobicity and polarity can be inferred from a continuous scale (*e.g.* Meiler et al., 2001). An alternative way to describe the similarity between their functions is to cluster the 20 amino acids into seven physicochemical groups (Figure 1.4). Aliphatic residues are relatively small (*e.g.* Glycine and Alanine) and do not contain aromatic, hydroxyl or charged groups in their side chains. Aromatic residues typically have the ability to form π-stacking using their aromatic ring. Basic and Acid groups are characterised by their charged amine groups (=NH^+-,=NH_2^+,-NH_3^+) and carboxylic acid (-C(=O)OH)

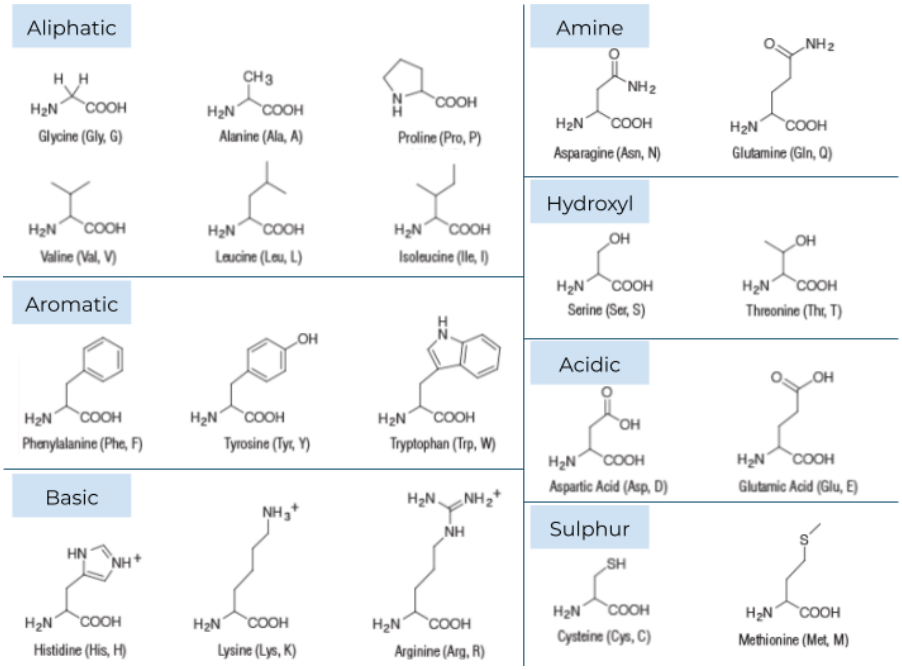


Figure 1.4: The chemical structures of amino acids, grouped by their physicochemical properties.

groups respectively. Residues in the "Amine" group have an amide (-C(=O)NH_2), and those in the "Hydroxyl" group all have a hydroxyl (-OH) group in their side chains. Finally, the "Sulphur" group has a sulphur atom in their side chain enabling the formation of disulphide bridges.

The residue composition in the paratope collectively determines the physicochemical features of the paratope and how it interacts with the epitope (*e.g.* MacCallum et al., 1996; Ramaraj et al., 2012; Qiu et al., 2015; Nguyen et al., 2017). Aromatic residues are enriched in paratopes compared to general protein-protein interfaces (Ramaraj et al., 2012; Kringelum et al., 2013). Ramaraj et al. (2012) quantified the residue "substitutability" and found that paratopes are inherently more specific than general protein surfaces. Both of these observations point towards a distinction between antibody paratopes and general protein-protein interfaces.

The intermolecular interactions at the antibody-antigen interface have also been studied. Qiu et al. (2015) developed chemical fingerprints to describe the residue

types in the paratope involved in interactions. Nguyen et al. (2017) observed the prominence of water bridges at the antibody-antigen interfaces. These findings suggest that both the physicochemical properties and interaction patterns in the antibody paratopes are important to the complementary binding with the epitopes.

1.2.2 T-cell receptors (TCRs)

TCRs are found on the surface of T cells and recognise peptide fragments from foreign molecules. They bind to peptides presented by the MHC, and their binding is known to be polyspecific (Rudolph et al., 2006). There are two types of TCR heterodimers: $\alpha\beta$TCR and $\gamma\delta$TCR (Janeway et al., 2001). In humans, most TCRs are $\alpha\beta$TCRs (Figure 1.1). Similar to antibodies, their α and β chains are formed by the somatic rearrangement of the respective V, D and J genes of the TCR loci. The random combination of these genes, alongside further diversification mechanisms (*e.g.* junctional diversity), can produce great variations of up to millions of unique TCRs (Attaf et al., 2015). TCRs, unlike antibodies, do not undergo SHM (Janeway et al., 2001). They rely only on gene rearrangement and junctional diversity to attain CDR variability.

TCRα chain is made from the V and J genes, while TCRβ chain is assembled from the V, D and J genes (Schroeder and Cavacini, 2010; Attaf et al., 2015). As in antibodies, the sequence and structural diversity of TCRs is concentrated in the six CDRs: three in the TCRα chain ($\alpha1$ – $\alpha3$) and three in the TCRβ chain ($\beta1$ – $\beta3$). These six CDRs make up the majority of the antigen binding site of the TCR.

Structurally, TCR heterodimers share a similar β-sandwich architecture to that seen in antibodies (Figure 1.1; Dunbar et al., 2014a). The six CDRs are responsible for engaging different parts of the peptide-MHC complex. CDRs 1 and 2 on both the α and β chains typically contact the MHC's conserved α-helices (Blevins et al., 2016; Sharon et al., 2016), while the two CDR3's almost always contact the peptide

antigen (Cole et al., 2014; Glanville et al., 2017). The structural complementarity between the binding site of a TCR and its cognate antigen governs the binding interactions. As the CDRs form the majority of the binding site, their conformations are critical to the binding.

1.2.2.1 TCR structural characteristics

There are far fewer TCR structures than antibody structures available (∼440 PDB structures containing TCRs and ∼4200 for antibodies as of September 2020; Dunbar et al., 2014b; Leem et al., 2018a). This lack of data has limited the number of cross-TCR structural studies. As noted in Section 1.2.1.3, the canonical form model has been widely studied in antibody structures, but only two studies have so far applied the concept to TCR CDRs (Al-Lazikani et al., 2000; Klausen et al., 2015). In the two studies, only 7 and 116 structures were used to yield 4 and 38 canonical forms respectively (Al-Lazikani et al., 2000; Klausen et al., 2015). In the non-canonical CDRβ3, similar dynamic studies to those carried out in antibodies have been done to study their flexibility and conformational selection upon polyspecific binding (Holland et al., 2018; Crooks et al., 2018; Fernández-Quintero et al., 2020b). We will discuss TCR canonical forms and their flexibility further in Chapter 3.

The limited structural coverage necessitates structural modelling to be able to expand the structural studies of TCRs. TCR homology modelling approaches were typically adapted from that in antibodies. LYRA calculates the BLOSUM similarity when selecting templates (Klausen et al., 2015). TCRModel considers germline and sequence similarity, with an option to refine CDR conformers with energy functions (Gowthaman and Pierce, 2018). Repertoire Builder uses multiple sequence alignment to select templates for modelling (Schritt et al., 2019). All of these tools have a similar accuracy of ∼1Å for CDRs 1 and 2, and ∼2.5Å for CDR3 loops on both α and β chains (see Chapter 3). Some were extended to model how the query TCR engages with the MHC by template modelling (Pierce and

Weng, 2013; Hoffmann et al., 2018; Jensen et al., 2019). These tools provide a single snapshot of the potential TCR conformations and do not consider any potential conformational selection or binding site flexibility.

1.2.3 Similarity and differences between antibody and TCR

As described in the above sections, antibodies and TCRs are both formed of the repertoire V, D and J genes from the respective antibody and TCR loci, and share similar β-sandwich folds (Dunbar et al., 2014a). More specifically, antibody light and TCRα chains are encoded by V and J genes, while antibody heavy and TCRβ chains are formed of V, D and J genes. These will therefore be considered as analogous pairs for all comparative analyses later in the thesis.

Various studies have commented on the binding site flexibility in these two types of proteins. Affinity-matured antibodies were observed to be more rigid (Mishra and Mariuzza, 2018). MD simulations and energy calculations revealed the flexibility in the TCR binding sites (Garcia et al., 1998; Morris and Allen, 2012; Rossjohn et al., 2015; Fernández-Quintero et al., 2020b). This could be the differentiating feature for the polyspecificity in TCRs but high specificity in antibodies.

1.3 Sequencing immune repertoire

Co-crystal structures contain valuable information for binding site analysis. The distances between the antibody and antigen residues and how these residues and atoms are configured with respect to each other are used to describe the region of the paratope and epitope and infer their interactions (Krawczyk et al., 2013, 2014; Stave and Lindpaintner, 2013; Jubb et al., 2017). As of September 2020, ∼4200 antibody structures have been deposited in the Protein Data Bank (PDB; Berman et al., 2000), of which ∼3000 structures are complexed with their antigens (Dunbar et al., 2014b; retrieved September 2020). However, our current structural data

suffers from two problems: a strong bias towards engineered antibodies and technical constraints in crystallising long loops (Dunbar et al., 2014b). Around 2700 PDB structures containing antibodies have been marked as "engineered" (Dunbar et al., 2014b), that implies only a small snapshot of the natural diversity of antibodies has been captured. Moreover, the constraints in crystallography techniques mean that antibodies with longer (and potentially flexible) loops are not well represented in the PDB (Marks and Deane, 2017). The limited structural coverage of the possible immune repertoire prompts the need to seek an alternative data source.

A possible alternative source of antibody data is the large set of antibody sequences generated by the next generation sequencing of the immunoglobulin repertoire (Ig-seq). These antibody sequences cover a greater breadth of antibody variability. There are several repertoires of this data in the Observed Antibody Space (OAS): almost two billion unpaired antibody sequences have been collected and collated (Kovaltsuk et al., 2018). With the rise of single-cell sequencing, an increasing number of natively-paired antibody sequences has also been curated (∼100,000 paired sequences in OAS as of September 2020; Kovaltsuk et al., 2018). However, without structural information, sequence data alone does not provide insights into antibody-antigen engagement. Methods to quickly annotate sequencing datasets with structural information are desired to act as a first filter, before carrying out detailed modelling procedures on the paired antibodies of interest (DeKosky et al., 2016; Krawczyk et al., 2018; Kovaltsuk et al., 2020).

1.3.1 Challenges in characterising antibodies in large antibody sequencing datasets

Sequence-based techniques have been used to characterise the repertoire dynamics during disease progression or vaccine response (*e.g.* Galson et al., 2015; Trück et al., 2015; Soto et al., 2019). These techniques are based on the premise that B-cells from the same clone, those with the same V and J genes and with highly similar CDRH3

sequences, will bind to the same epitope (*e.g.* Scheid et al., 2011; Hsiao et al., 2020). Whilst this "clonotype" analysis can identify antibodies with similar binding modes, intra-clonal diversity and inter-clonal convergence have been observed (Krause et al., 2011). For example, anti-lysozyme antibodies with different CDRH3 sequences and genotypes have been found to engage the same epitope (Pons et al., 2002). Sequence identity and germline origin alone may not fully reflect the binding modes of antibodies.

Alternative approaches are now appearing that involve combining *in vitro* screening and deep learning algorithms. Mason et al. (2019) presented a high-throughput algorithm that can predict on a 10^8 *in silico* screening library which sequence will bind to a desired target. However, the method required 10^4 antibody sequences with assayed target antigen specificity as a training set. The initial training set generation is highly resource-intensive. Therefore, the method is unable to generalise to other targets.

Adding structural information can better inform the potential binding site structures captured by the sequences. To capture the three-dimensional configuration of paratopes, computational antibody modelling tools (*e.g.* Sircar et al., 2009; Maier and Labute, 2014; Biasini et al., 2014; Almagro et al., 2014; Klausen et al., 2015; Leem et al., 2016; Li et al., 2019) have been developed to predict antibody structures from their sequences. These structural models can contribute to paratope analysis and give insights into possible interactions with the antigen (Kovaltsuk et al., 2017). A key drawback of modelling approaches is their scalability: one of the fastest tools takes ∼30s per antibody sequence (Leem et al., 2016). Modelling the two billion antibody sequences available in the Observed Antibody Space (Kovaltsuk et al., 2018) on 100 cores would take over 19 years to complete. A faster way to gain structural insights across the large amount of sequences is required to unravel the paratope shapes and physicochemical features.

1.3.2 Enriching sequence repertoires with structural information

The ever increasing volume of sequencing data has the potential to add to our understanding of antibody binding. Various studies have demonstrated how to enrich sequencing datasets with binding site structure annotation (DeKosky et al., 2016; Krawczyk et al., 2018; Kovaltsuk et al., 2020; Richardson et al., 2020), which in turn could aid in understanding antibody binding.

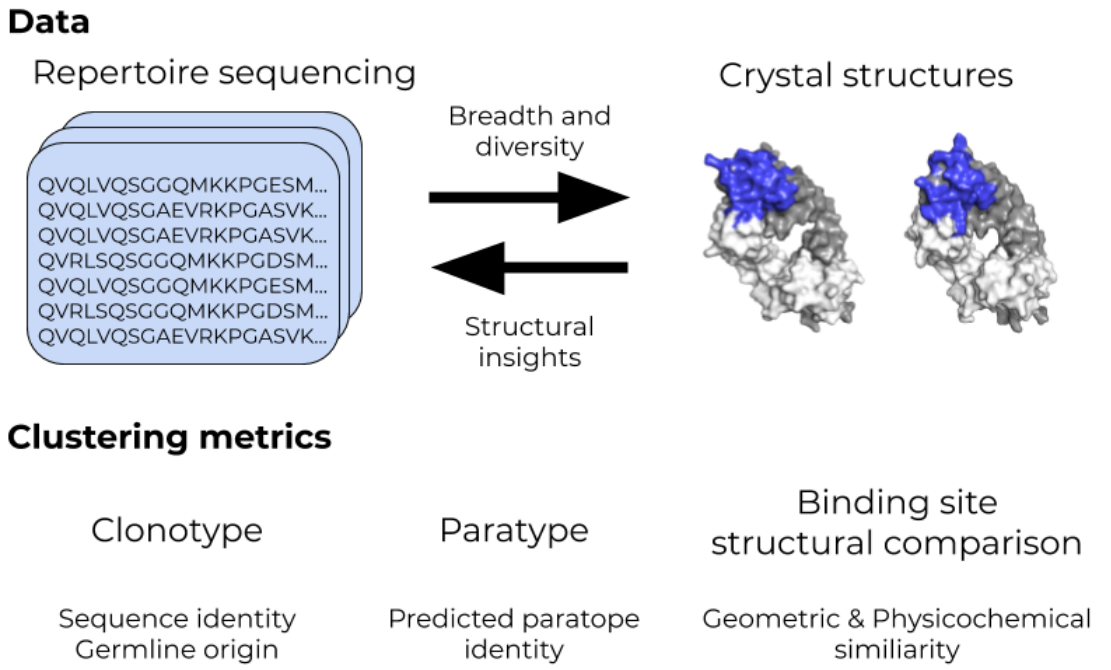


Figure 1.5: Datasets and clustering metrics for analysing immune repertoires and antibody structures. Data from repertoire sequencing and crystal structures can improve our understanding of antibody variability and binding site features. Important features are extracted by the clustering metrics: clonotype is a sequence-based metric that considers the sequence identity and germline origin; binding site structural comparison captures the geometric and physicochemical similarity between two surfaces; and a hybrid method "paratype" calculates the predicted paratope identity between two antibodies.

Figure 1.5 summarises the available data formats and the main approaches employed to study antibody variability and binding profiles. Sequence-based approaches tend to consider the sequence identity and germline origins alone,

for example in clonotype analysis. These methods are fast and scalable, but may miss structurally similar paratopes that differ in sequence (see Section 1.3.1). At the opposite end of the spectrum, comparing binding site structures often revolves around geometric and physicochemical similarity of the surface patches (see Section 1.4.2.2). Structure-based approaches face a bottleneck in the scalability of the homology modelling in the context of large Ig-seq datasets.

Hybrid methods use rapid prediction of structural features and/or the binding site from sequence. An example of such a method is paratyping that, instead of only considering the sequence identity and germline, clusters antibodies using the predicted binding site positions (Richardson et al., 2020). It is able to find highly sequence-dissimilar antibodies that bind to the same epitope, and is more scalable than structural modelling.

1.4 Antibody binding sites characterisation

As described in Section 1.2.1.2, the features of an antibody binding site determine its binding behaviour. The highest resolution method for studying antibody-antigen binding configurations is co-crystal complex structures. These give atomic level information but co-crystal complex structures are expensive and difficult to obtain (Dunbar et al., 2014b). In this section, we discuss the *in vitro* and *in silico* techniques that infer antibody binding at different levels of resolutions (Figure 1.6).

1.4.1 *In vitro* assays

1.4.1.1 Competition assay / epitope binning

Competition assays exploit the cross-blocking effect of antibodies that displace one another if they bind to similar or neighbouring epitopes (Kwak and Yoon, 1996; Abdiche et al., 2017). One of the most popular epitope binning methods is the array surface plasmon resonance (SPR). In epitope binning array SPR, a classical

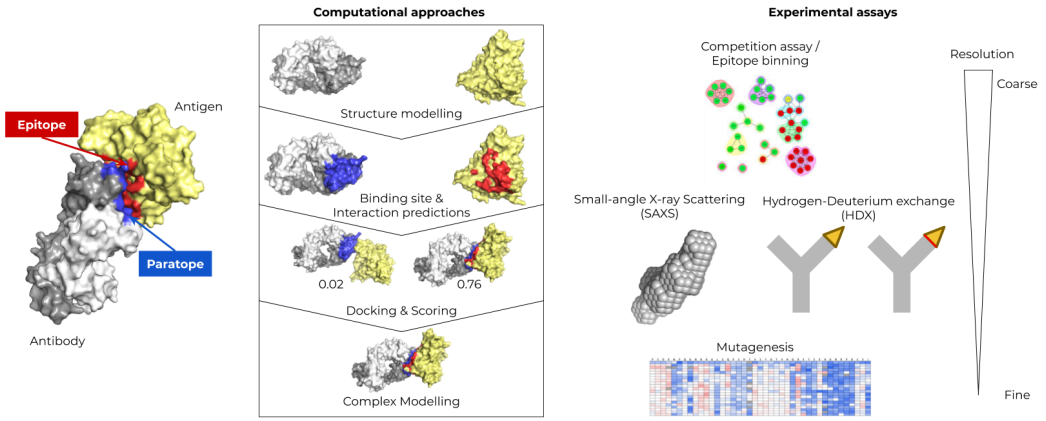


Figure 1.6: Binding site analysis techniques. (Left) Definition of antibody (heavy chain in grey, light chain in white), antigen (in yellow), paratope (in blue) and epitope (in red). Paratope on the antibody and epitope on the antigen are considered as residues within 4.5Å of the binding partner. (Middle) Computational approaches to study binding mode. The global antibody and antigen structures or models are used as starting points. Their binding sites and potential interactions are predicted, to constrain docking and pose generation. These poses are then scored for complementarity. The top-rank docked pose is selected as the modelled complex. (Right) Experimental binding assays at varying levels of resolutions. Competition assays inform the epitopic specificity of the antibodies. Small-angle X-ray Scattering (SAXS) outlines the overall shape of the complex and Hydrogen-Deuterium exchange (HDX) coupled with mass spectrometry describes upto peptide-level resolution of the binding regions. Mutagenesis identifies the per-residue contribution to binding.

"sandwich" format is used: the first specific antibody ("ligand") is immobilised on a sensor chip and captures its specific antigen, while the second antibody ("analyte") flows on the sensor chip. If the two antibodies compete with each other, or displaces one another, a reduced binding is detected and they are considered "blocked" (Abdiche et al., 2017). Binding events change the refractive index and resonance angle of the base surface film. This is detected and calculated as the variation in response units over time. These quantities can be translated back to the binding kinetics of the two ligand and analyte antibodies (Malmborg and Borrebaeck, 1995). In epitope binning studies, pairwise comparisons between all antibodies in a set are used to create a competition profile, from which clusters of antibodies with similar blocking behaviours are identified. This method gives a coarse representation of which binders may share similar target sites, as minimal

epitope overlap (or even neighbouring epitopes) can be sufficient for a pair of antibodies to compete with each other (Abdiche et al., 2017). No binding mode information can be extracted directly from the competition matrix.

1.4.1.2 Small-angle X-ray scattering (SAXS)

Small-angle X-ray scattering (SAXS) informs the global molecular shape of protein complexes. Unlike the gold-standard X-ray crystallography, a crystalline sample is not needed; SAXS can capture flexible molecules in solution (Sutton et al., 2018). In contrast to nuclear magnetic resonance that does not generally tolerate large molecules, the sensitivity of SAXS increases with the molecular size (Göbl et al., 2014). Even though SAXS is unable to determine the exact atomistic coordinates, it has been used in conjunction with molecular modelling techniques to study antibody binding (Castellanos et al., 2017; Sutton et al., 2018).

1.4.1.3 Hydrogen–deuterium exchange mass spectrometry (HDX-MS)

A more refined approach is hydrogen–deuterium exchange mass spectrometry (HDX-MS). This technique measures the solvent accessibility of the antibody or antigen surface. The backbone amide hydrogen of the peptide is displaced by deuterium. If the stretch of peptide is at the binding interface, the rate of exchange is reduced (Sevy et al., 2013). It can also reflect the allosteric effects upon binding. A key advantage of HDX-MS is that it requires relatively low amounts of proteins for the analysis, which is desirable at early-stage antibody discovery (Puchades et al., 2019). The mainstream protocols only allow up to the range of peptides in the immediate proximity of the binding site (*e.g.* Zhang et al., 2018; Puchades et al., 2019). However, a more recent version incorporating electron transfer dissociation was shown to improve the resolution to reach residue-level refinement of the binding interface (*e.g.* Pan et al., 2015; Huang et al., 2020).

1.4.1.4 Mutagenesis

To achieve residue-level resolution, point mutations of the interacting proteins have been used to highlight key binding residues. Mutagenesis studies measure the binding kinetics upon mutation of a single or a set of residues (*e.g.* Greaney et al., 2020). Two main types of mutations are typically introduced: a site-specific mutation to alanine ("alanine scanning") or a complete mutagenesis covering mutations to all of the 20 amino acids at the selected positions. Mutants are expressed in yeast or other bacterial or mammalian systems. Single, double or triple mutants can be made following the protocol. Misfolded proteins are removed by flow cytometric sorting (Cherf and Cochran, 2015), and the binding kinetics of the antibody to the mutant antigen are measured by SPR or other biosensing platforms (Greaney et al., 2020). This method unravels the binding contribution of each position in the paratopes and epitopes. However, structural integrity may be compromised by the mutations, leading to spurious results (Abbott et al., 2014).

1.4.2 Computational techniques

As mentioned in Section 1.2.1.4, antibody-antigen binding interfaces have different compositions to that of general protein-protein interfaces. It is important to identify the specific binding residues in the paratope and epitope to aid interface design. Figure 1.6 shows a common computational pipeline of binding site analysis. Following the structural prediction of both the antibody and antigen of interest, their binding sites (paratopes and epitopes) and potential interactions are predicted and mapped onto the model structures. These two binding sites are used to guide complex modelling that involves selecting a top-ranked docked pose. Since macromolecular docking has been as challenge (Lensink et al., 2019), alternative avenues have been explored and borrowed from small molecular-protein binding. An example of this is binding site comparison. Below, we focus on the antibody and discuss the existing methods for paratope prediction and surface comparison.

1.4.2.1 Paratope prediction

Paratope features have been widely studied to pinpoint which residues on the antibody are likely to contact the antigen. As discussed in Section 1.2.1.2, the majority of the paratope is covered by the CDR loops. Across different definitions of the CDRs, these loops cover 80% of the antigen-contacting residues on average. The additional antigen-binding regions are in the framework, especially between CDRs 2 and 3 as they are often situated in the proximity of the antigen (Kunik et al., 2012a,b). On the other hand, the CDR loops may contain residues that do not participate in binding. To accurately predict paratopes, a number of sequence-based and structure-based methods have been proposed to annotate Ig-seq datasets and will be discussed below.

Sequence-based paratope prediction methods have been built upon structural paratope features and their sequence motifs. Paratome aligns the input sequence or structure to the existing antibodies and selects the structurally conserved regions as the paratope (Kunik et al., 2012a). Liberis et al. (2018) recently built Parapred, that uses recurrent and convolutional layers to process a curated set of antibody CDR sequences by their physicochemical properties, aggregating different CDR types for training and returning the binding probability of each CDR residue in an antibody. Parapred reported its performance by the area under the Receiver Operating Characteristic (AUROC) curve of 0.878±0.004, over the CDR regions of the antibody only. Taking a further step to predict the types of intermolecular interactions that a predicted paratope residue would participate in, proABC and proABC-2 were developed using a random forest and a convolutional neural network respectively (Olimpieri et al., 2013; Ambrosetti et al., 2020). Compared to Parapred using the Parapred's training set, proABC-2 has a reported AUROC of 0.91 in paratope prediction that is slightly superior to Parapred. The authors also used the predicted interactions to inform antibody-antigen complex modelling. These sequence-based approaches only require the antibody sequence as the input, and their fast run-time has enabled the annotation of large sequencing datasets (Richardson

et al., 2020).

Structures can inform the surface exposure of the residues and how they are positioned with respect to each other. Graphlet-based comparison (Krawczyk et al., 2013), Zernike descriptors (Daberdaku and Ferrari, 2019) and geometric convolutional neural network (Pittala and Bailey-Kellogg, 2020) leverage the spatial and physicochemical features to identify paratopes on an input antibody structures. The three tools have AUROC's of 0.84, 0.95 and 0.95 respectively in the evaluation of Pittala and Bailey-Kellogg (2020). In the context of analysing the paratope diversity in sequencing datasets, model structures would need to be built. This would be a limiting factor as the current structural modelling tools are not yet scalable.

As an antibody may use different paratope residues to bind to different antigens (Bostrom et al., 2009; Lee et al., 2017), an additional and important feature for paratope prediction is the antigen information. Partner-aware paratope prediction considers the cognate antigen and suggests paratope residues on the input antibody sequences or structures. Antibody i-Patch is a structure-based method that uses the structures of both the antibody and antigen to compute the binding likelihood of residues on the antibody (Krawczyk et al., 2013). With sequence information alone, Deac et al. (2019) extends the original Parapred model to incorporate antigen sequence information through attention layers in the neural network architecture. As both of these tools consider the properties of the cognate antigen, they should in theory be able to make bespoke predictions for antibody-antigen pairs.

All of these paratope predictions tools were trained on structural datasets. However, to apply them on Ig-seq datasets, we need to address the challenges outlined in Section 1.3.1. It is necessary to validate these on sequencing datasets that could potentially cover a more diverse "paratope space", for example the larger

percentage of longer CDRH3 than seen in crystal structures (Kovaltsuk et al., 2020).

1.4.2.2 Searching for similar binding sites with surface descriptors

The surface shape of a paratope can determine the type of its cognate antigen. Lee et al. (2006) showed that small molecule and peptide antigens prefer a deep pocket in the paratope, whereas most of the paratopes for a protein antigen adopt a "planar" shape. Similar messages were brought about by applying 3D Zernike descriptors to antibody-antigen binding site structures (Di Rienzo et al., 2017). More detailed descriptors have also been applied to compare paratope shapes. Below we included six major families of surface descriptors that have been used to capture the surface shape and chemical features of protein binding sites (Figure 1.7).

Tessellation

Surface tessellation is built upon triangulation of the surfaces, whereby triangles are created from all of the atomic coordinates on the protein surface (Figure 1.7A). The surfaces are then aligned and the overlap of triangles between two surfaces is used as a similarity measure (Liang et al., 1998; Xie and Bourne, 2008; Schalon et al., 2008; Albou et al., 2009; Zhou and Yan, 2014). It can also be used to compare the depth of two cavities (Albou et al., 2009).

Surface curvature

Fractal atomic density, *e.g.* Surfactal (Kuhn et al., 1992), measures the surface curvature by the gradient of atomic density. The atomic density of a point is the number of surrounding atoms within a defined radius. For instance, in Figure 1.7B, the red circle encloses four atoms, while the blue circle only includes one atom. The atomic density is higher around surface points which are deeper in the grooves (red circle), and lower around surface points which are more exposed (blue circle). By

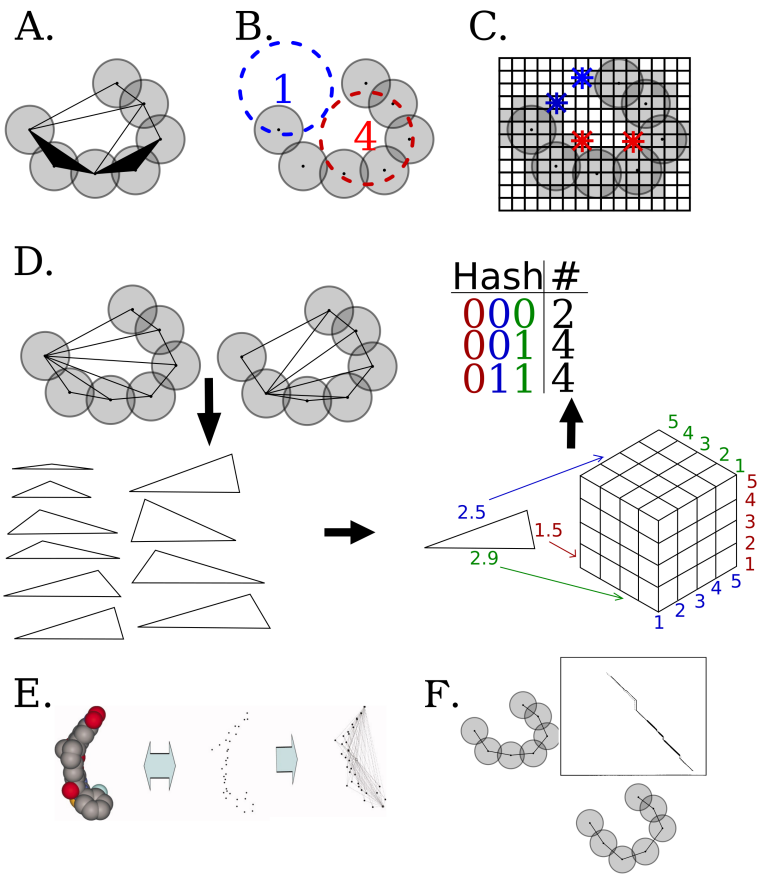


Figure 1.7: Shape descriptors. **A.** Tessellation. The surface is triangulated from all the atomic coordinates. Discrete flow analysis removes triangles with any edges that fall outside the surface region. **B.** Surface curvature. The atomic density of a point is reflected by the number of surrounding atoms within a defined radius. **C.** Grid-based methods. A grid is blocked if it is buried in the protein. A probe surveys its neighbouring grids when rolling over the surface, to vectorise the degree of buriedness. **D.** Geometric hashing. All possible triplet combinations of atoms in the small molecule or protein pocket are extracted. These triangles are hashed by their edge distances and vertex types, represented by the three-dimensional cube. The total number of triangles in each hash bin is tallied and recorded into a frequency table ("hash table"). Molecules with similar geometries should have comparable hash tables. **E.** Ultrafast Shape Recognition (USR; extracted from Ballester and Richards (2007)). USR is an alignment-free method that identified four reference points (explained in the chapter) on the molecule and computed the pairwise distances as vectors for comparison. **F.** Distance map-based algorithms. The inter-atomic distances of the binding site are used to build a distance matrix. To search for the best alignment between the two binding sites, dynamic programming algorithms optimise an objective function that considers structural and chemical matches.

this scheme, the atomic density is continuous across the surface. MacCallum et al. (1996) suggested a threshold for the definition of surface concavity based on this measure, which could categorise surfaces by their depths.

Grid-based methods

A grid-based approach can facilitate the calculation of the cavity opening, elongation and other shape descriptors. For instance, in PocketPicker (Weisel et al., 2007) and Vectorial Identification of Cavity Extents (VICE; Tripathi and Kellogg, 2010), a 3D-grid is placed to discretise the surface of the protein. A grid is blocked if it is buried in the protein structure; otherwise it is exposed. A probe is rolled over the surface of the protein to survey its environment to calculate the percentage, and vectorise the positions of buried neighbouring grids (Figure 1.7C). In the example shown in Figure 1.7C, the red probes hit four and six blocked grids in the surrounding eight grids, which implies that they are positioned in a buried cavity. The blue probes only hit two and three blocked grids, suggesting that they are situated on a more exposed region. The set of vectors gives an indication of cavity dimension and shape including the width of opening, elongation and sphericity.

Geometric hashing

Geometric hashing has been widely used in computer vision to compare similar objects. It involves identifying reference points on an image, constructing triangles from all possible triplets of points set, and encoding the information of the triangles into a hash table. By comparing the hash tables of two images, a similarity score suggests how likely the two images contain the same object.

In the application of geometric hashing to small molecule binding, the spatial coordinates and chemical features (*e.g.* by element or atom types) are encoded (Figure 1.7D). The pharmacophores (atoms that participate in interactions) in a

small molecule are used to build "triangles". All triplet combinations are hashed by the edge features: the length of the edge that is put into discretised bins of distances, and the vertices' chemical types at the two ends of the edge. A frequency table of the hashes ("hash table") is constructed for each molecule. The more overlap there is between the hash tables of two molecules, the more similar the molecules are. Previous methods have shown success in identifying geometrically similar surfaces in protein-ligand and protein-protein applications (Wallace et al., 1997; Shulman-Peleg et al., 2004, 2005; Wood et al., 2012; Núñez-Vivanco et al., 2016).

More recently, Ligity (Ebejer et al., 2019) extends to consider tetrahedrons of pharmacophore quadruplets in protein-ligand structures. With the fourth points, this method is able to consider the chirality in the small molecule conformers by calculating the volume of the tetrahedrons and encoding chirality as a feature in the hash function. Ligity was used to screen for conformers which are similar to the ligand in the query protein-ligand complex structures, and showed consistent results with docking but was much faster.

These methods have shown promises in small molecule characterisation and discovery. However, when they were adapted for protein-protein interfaces, the larger protein-protein interfaces compared to the protein-ligand pockets substantially increased the algorithm run-time (Shulman-Peleg et al., 2004, 2005). The all-atom comparison between protein surfaces becomes unscalable to datasets in the size of Ig-seq datasets. More details will be covered in Chapter 4.

Ultrafast Shape Recognition (USR)

USR (Ballester and Richards, 2007) was introduced as an alignment-free method to compare the geometry of small molecules (Figure 1.7E). Four reference points are identified in a small molecule based on the molecular centroid and three extrema

of the molecule: (a) the closest atom to the molecular centroid; (b) the farthest atom to the molecular centroid; and (c) the atom farthest from (b). Selecting the same reference points based on the extrema on every structure avoids the need to align the structures during the structural comparison. Instead, USR uses the vectors corresponding to each reference points to compute their structural similarity. The pairwise distances between each atom and these reference points are calculated and the list of distances are stored for each reference point. The first (mean), second (variance) and third (skewness) statistical moments of the distances with respect to each of the reference points are encoded in 12 descriptors (three moments for each of the four reference points). By comparing the vectors of descriptors, similar surfaces should have a small numerical difference in the vectors. Extensions of the USR method incorporated physicochemical features of the atoms such as electrostatics in Electroshape (Armstrong et al., 2010) and CREDO atom types in USRCAT (Schreyer and Blundell, 2012) as additional vectors. These extensions provided means for describing the topology and chemical features, but their applications on protein-protein interfaces has not been evaluated yet.

Distance map-based algorithms

Distance maps have been extensively used in protein structural alignment, for example in combinatorial extension (Shindyalov and Bourne, 1998) and Dali (Holm and Sander, 1995). Recently, they have been applied more locally to quantify the similarity between binding sites, independent of sequential orders of the residues (Figure 1.7F). For example, Gao and Skolnick (2013) and Mirabello and Wallner (2018) developed different objective functions and applied stochastic approaches to speed up the optimisation process. APoc (Gao and Skolnick, 2013) uses the Linear Sum Assignment Problem to optimise for the "Pocket similarity score", calculated from the backbone geometry, side chain orientation and chemical similarity. InterComp (Mirabello and Wallner, 2018) considers the distance and BLOSUM substitution between the aligned residues as their objective function for

simulated annealing. These methods have shown promising results in identifying similar protein surfaces. However, the main bottleneck is the inherently long computational time of dynamic programming methods.

1.5 Thesis overview

In this work, I have analysed and predicted properties of antibody and TCR binding site structures. Inspired by initial structural analysis, we describe the development of descriptors optimised for application on BCR sequencing datasets, to explore the binding site structure diversity captured by immune repertoires.

1.5.1 Chapter 2

In this chapter, we focus on the CDR canonical forms in antibodies and propose a fast, sequence-based predictor for canonical forms annotation, SCALOP. We show that SCALOP makes consistent predictions to a database-search method on an Ig-seq set, but in a much shorter time. As more structural data becomes available, we see an increase in the number of canonical forms and implement an auto-updating database in SCALOP to ensure that it has captured the most recent information about the canonical forms. We further demonstrate its application on a mutagenesis set to show that SCALOP is able to rapidly highlight the binding site structural differences of a set of antibody sequences.

1.5.2 Chapter 3

This chapter covers the analysis of antibody and TCR binding site structures, in an attempt to decipher the mechanisms that differentiate their biological functions. We find that TCR CDRs can adopt multiple conformations, more often than their antibody counterparts. Building on the increased variability of TCR CDR conformations, we develop a multi-state TCR modelling tool, TCRBuilder, to

capture this phenomenon. We show that TCRBuilder is comparable to the existing tools, with the additional benefit of portraying the multi-state conformations of TCR binding sites.

1.5.3 Chapter 4

The majority of the sequence-similar antibodies adopt the same binding mode against their cognate antigens. However, sequence-dissimilar antibodies can also bind to the same epitope. Sequence-based metrics would have missed these cases, and a structural descriptor is needed. This chapter introduces Ab-Ligity, a structure-based metric that identifies paratopes against highly similar epitopes. We compare Ab-Ligity with an existing surface comparison tool built for general protein surfaces.

1.5.4 Chapter 5

Paratope analysis and prediction have been built upon available structural data that is biased towards structures with shorter loops due to technological constraints. To apply these tools to sequencing datasets that capture a more diverse space, we need to assess the applicability of these paratope prediction tools. This chapter covers the preliminary results of our paratope analysis and an appraisal of a deep learning paratope prediction tool, Parapred. We then incorporate some key paratope features into simpler, more interpretable machine learning models. To assess the length-dependence of the models, we observe their behaviours when trained on shorter loops and tested on longer loops, and compare to that of Parapred. Due to the interpretable nature, we further identify the features that the simpler models use to determine paratopes against protein or peptide targets.

1.5.5 Chapter 6

In this chapter, we conclude the findings of this DPhil thesis. Based on the results, we propose future directions that can be pursued to harness the data we currently have and fill the gaps to understand antibody and TCR binding.

2

Sequence-based Antibody Canonical Loop Structure Prediction

Contents

This chapter is adapted from my published work "SCALOP: sequence-based antibody canonical loop structure annotation" in *Bioinformatics* (Wong et al., 2019a).

2.1 Introduction

With the development of next generation sequencing, over 500 million antibody sequences are now available and collated in the public domain (Kovaltsuk et al., 2018). A rapid way of gaining structural insights across large sequencing datasets is necessary. As mentioned in Section 1.2.1.2, CDRs form the majority of the paratope and canonical forms have been observed in five of the six CDRs. These canonical forms are known to be predictable from sequence alone (*e.g.* Chothia and Lesk, 1987; Martin and Thornton, 1996; Nowak et al., 2016). Thus a canonical form predictor can give a rapid approximation of the paratope shape. A sequence-based

predictor of paratope shapes can be used to screen for possible binders with the preferred geometries against the target epitope.

Since the introduction of CDR canonical forms in 1987 (Chothia and Lesk, 1987), they have been revisited many times, but each update has been a static snapshot of the data available at that time point (see Section 1.2.1.3). By examining these renewals, we are able to illustrate how the growth of structural data has continuously modified our understanding of CDR loop structures, from the 10 canonical forms seen in 1987 (Chothia and Lesk, 1987) to the 26 in 2016 (Nowak et al., 2016). Based on these definitions of canonical forms, several sequence-based canonical form prediction methods have been developed (*e.g.* Chothia and Lesk, 1987; Martin and Thornton, 1996; Nowak et al., 2016). Chothia and Lesk (1987) suggested structurally-determining residues for canonical form assignment. Using a similar approach, Martin and Thornton (1996) published a freely available web server that takes as input complete and paired antibody sequences, but lacks a software package for bulk processing. Hidden Markov models have also been built for cluster assignment but are not readily available (North et al., 2011; Nowak et al., 2016). AbDesign within Rosetta uses position-specific scoring matrices (PSSMs) of the canonical forms for predicting the CDR backbone conformations (Lapidoth et al., 2015). However, none of these tools uses an auto-updating database, and none provides both a web interface and a freely available software package for large-scale sequence analysis.

Below we describe our fast, sequence-based antibody canonical loop structure prediction method, SCALOP, supplemented with an auto-updating database. This tool is publicly available at `http://opig.stats.ox.ac.uk/webapps/scalop`.

2.2 Method

The database of SCALOP is built following the definition of Nowak et al. (2016), and PSSMs are then constructed for each of the structural clusters when making a prediction. SCALOP takes one or a set of amino acid sequences of full antibody chains as input. It then numbers the sequence with ANARCI (Dunbar and Deane, 2016), and scores the extracted CDR sequences against PSSMs of its clusters. The input CDR sequence is then assigned to the cluster with the maximum score above a scoring threshold. The Protein Data Bank (PDB; Berman et al., 2000) code and chain identifier of the assigned cluster's median structure is returned along with the canonical class. The database is updated monthly. Below is a detailed description of the steps taken.

2.2.1 Length-independent clustering of CDR loop structures

We followed the protocol outlined by Nowak et al. (2016) to carry out length-independent clustering of the CDR structures.

2.2.1.1 CDR loop extraction from protein structures

All X-ray structures available in SAbDab (Dunbar et al., 2014b) as of 10$^{\text{th}}$ July 2017, with a resolution of $\leq$2.8Å, were used initially (hereafter the "SAbDab set"). For this work, we adopted the IMGT numbering scheme (Lefranc et al., 2015) and the CDR definition described by North et al. (2011). CDR loops with no missing residues and no B-factors of backbone (C, C_α, N and O) atoms $\geq$80 were considered (Nowak et al., 2016).

2.2.1.2 Cluster formation

Five residues before and five after the CDR termini were used as the anchors for structural alignment. The CDRs were superposed using only the 10 anchor residues and the pairwise backbone root-mean-square deviation (RMSD) between loop structures were calculated as follows:

$$\mathrm{RMSD}(\mathbf{v}, \mathbf{w}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}((v_{ix} - w_{ix})^2 + (v_{iy} - w_{iy})^2 + (v_{iz} - w_{iz})^2)}, \tag{2.1}$$

where $\mathbf{v}$ and $\mathbf{w}$ are two sets each of n atoms, and x, y and z refer to the x-, y- and z-coordinates of these atoms.

The RMSDs between all the atoms in the loop structures were used to form the cost matrix. For loop structures that differ in length, a dynamic time warping (DTW) algorithm was used to find the optimal structural alignment between the backbone atoms (in a method similar to Needleman-Wunsch sequence alignment algorithm; refer to Nowak et al., 2016 for detailed implementation). Density-based spatial clustering of applications with noise (DBSCAN; Ester et al., 1996) was used to carry out the structural clustering due to its ability to search for densely populated regions and group the neighbouring points, without knowing the number of expected clusters. Nowak et al. (2016) selected optimal distance cut-offs by the Ordering Points to Identify the Clustering Structure (OPTICS; Ankerst et al., 1999) algorithm. The clustering thresholds used in this study are the same as those in Nowak et al. (2016), except for L2 where a clustering threshold of 1Å was used. This led to three H1, four H2, twelve L1, one L2 and seven L3 clusters. The summary statistics of the clusters are shown in Appendix A.1.

2.2.2 Building the position-specific scoring matrix

2.2.2.1 Constructing PSSMs for the clusters

Consistent with the description in Nowak et al. (2016), we defined a cluster for use in prediction as a set of CDR structures with at least six unique sequences. A

cluster is named as follows: the first two letters represent the type of CDR (H1 or H2 *etc.*), followed by the sequence lengths found in the cluster, and completed with an alphabet representing the rank of the cluster in descending sizes. For instance, an L3 cluster which contains length-10 and length-11 sequences (10,11), and has the second highest number of unique sequences among all clusters which contain both length-10 and length-11 sequences (B), is called "L3-10,11-B".

We constructed a PSSM for each cluster based on the frequency of amino acids found at a given position within the cluster:

$$M_{k,j} = \log_2 \frac{p_{k,j}}{b_k}, \tag{2.2}$$

where $M_{k,j}$ is the element score and $p_{k,j}$ is the probability of observing the amino acid k at the IMGT-numbered position j in the cluster, and b_k is the background probability of amino acid k, which is considered to be the same for all amino acid types (*i.e.* 0.05; alternative background probability derived from the distribution of amino acids in antibodies could also be used, but is not considered in this study for simplicity). A pseudo-count of 0.001 was added to all elements with no observations to prevent computational errors.

2.2.2.2 Scoring the PSSM and making a prediction

To make a prediction, we only considered clusters that contain members of the same sequence length as the target sequence. For instance, a length-10 L3 sequence will be considered against PSSMs of L3-9,10-A, L3-10-A, L3-10-B and L3-10,11-A. The PSSM score for a target sequence for cluster c is s_c given by:

$$s_c = \sum_{j=J_0}^{J} M_{k,j}, \tag{2.3}$$

where J is the set of positions in the target sequence. If the highest total score is above an assignment threshold (see Section 2.2.3 and Appendix A.2), an assignment is made to the cluster with the highest total score.

Nearly all of the L2 loops (2741/2765) in the SAbDab set are of length eight and we observed that 99.3% (2721/2741) of these length-8 L2 loops are clustered

in our L2-8-A. Henceforth, all length-8 L2 loops were assigned to a single cluster; loops of other lengths were not assigned to any clusters as we did not have any clusters of other lengths. This resulted in the same precision and recall as the selected threshold (see Table 2.1).

Table 2.1: Recall, precision and coverage at the selected thresholds.

CDR (threshold)	Recall (%)	Precision (%)	Coverage (%)
H1 (0.5)	99.45	89.26	93.75
H2 (-1.5)	99.89	93.6	97.54
L1 (-0.5)	99.84	95.67	97.38
L2 (-1)	100	99.13	98.5
L3 (0)	99.26	93.31	91.69

2.2.3 Cross-validation for threshold selection

We carried out leave-one-out cross-validation on the SAbDab set. Within a cluster, only unique sequences were used. For non-clustered sequences, only unique sequences in the set were retained for cross-validation.

For each loop, if the backbone RMSD between the actual structure and any members of the assigned cluster was <1.5Å, this was labelled a true positive (TP); otherwise this was labelled a false positive (FP). If the loop was not in any cluster and was not assigned to any cluster, this was labelled a true negative (TN). If the loop was in a cluster but was not assigned to any cluster, this was labelled a false negative (FN).

We used the following definitions for the calculation of recall, precision and coverage:

$$Recall = \frac{TP}{TP + FN}, \tag{2.4}$$

$$Precision = \frac{TP}{TP + FP}, \tag{2.5}$$

$$Coverage = \frac{TP + FP}{TP + TN + FP + FN}. \tag{2.6}$$

For each CDR, we calculated the precision and recall for different assignment thresholds. We then calculated the F_1 score for each threshold:

$$F_\beta = (1 + \beta^2)\frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall}, \tag{2.7}$$

where $\beta = 1$.

The maximum total scores were between -4 and 4 (Figure 2.1). We performed a parameter sweep between these values using an increment of 0.5.

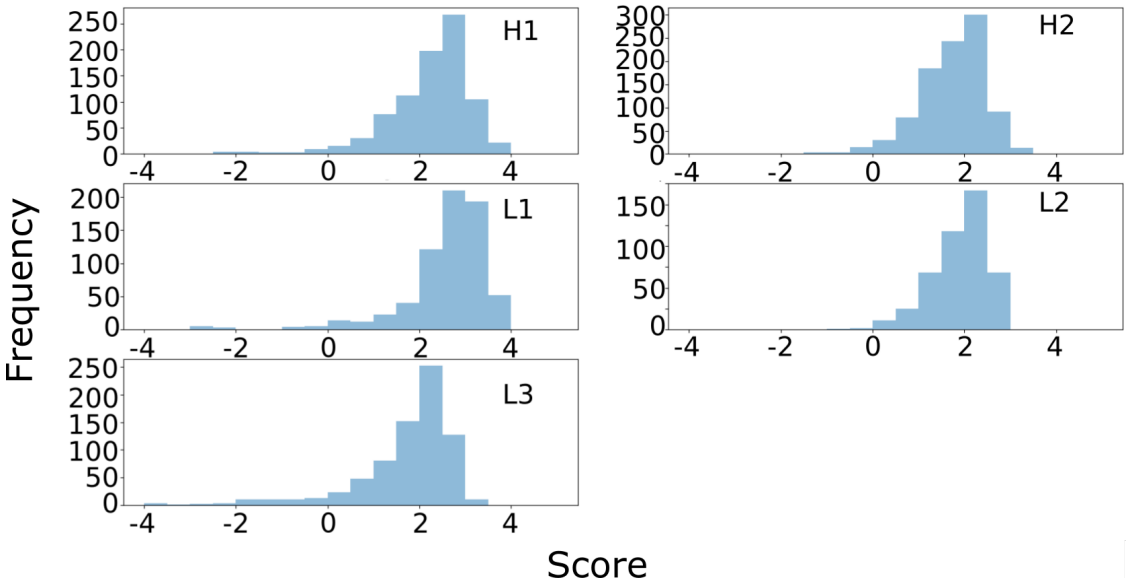


Figure 2.1: Maximum total score distributions for different CDR loops during cross-validation.

We selected the scoring threshold with the highest F1 score (Appendix Table A.3). Table 2.1 shows the corresponding recall, precision and coverage of the selected threshold for each CDR type.

2.2.4 Benchmark with FREAD

FREAD is a database-search method for loop structure prediction (Deane and Blundell, 2001; Choi and Deane, 2010). The algorithm extracts fragments in loop structures to build its database. To query FREAD, it takes as input a sequence and the anchor separation in the structure that the query sequence is to be grafted on. The environment substitution score (ESS) is calculated by the dihedral angle ranges

and amino acid substitution patterns in the query and database fragment structures.

In this study, FREAD was run using only structures whose PDB code and chain identifier are found in the SAbDab set. As in SAAB (Krawczyk et al., 2018), we selected the antibody framework structure onto which the query loop was grafted, by sequence identity to the native framework. To select a prediction, we used length-dependent environment substitution score (ESS) cut-offs (Krawczyk et al., 2018; Kovaltsuk et al., 2020):

- Lengths < 13: 25
- Lengths 13-16: 40
- Lengths > 16: 55.

As with the standard FREAD protocol, the decoy with the top ESS above the length-dependent ESS cut-off, and the lowest anchor RMSD with the model framework was selected as the FREAD prediction. FREAD does not make a prediction if none of the decoys are above the corresponding ESS cut-off, or when the decoy has an anchor RMSD of $\geq$1.0Å.

2.2.4.1 Coverage and precision of FREAD on loops in the SAbDab set

We ran a leave-one-out cross-validation on all structures within the SAbDab set. For each case, the frameworks and CDR loops from identical antibody sequences were eliminated. The same measures of correctness were used for FREAD as for SCALOP. A true positive prediction refers to a case where the backbone RMSD between the actual and predicted structures was <1.5Å; otherwise it was a false positive. If the minimal backbone RMSD between the actual structure and any loop structures was $\geq$1.5Å, it was considered a true negative if FREAD does not make a prediction; otherwise the lack of a FREAD prediction was considered a false negative. The calculation for the coverage and precision are the same as in Section 2.2.3.

2.2.5 Predicting on the Immunoglobulin gene sequencing (Ig-seq) set

To assess the speed and portion of consistent predictions made by SCALOP and FREAD, we ran both predictors on a set of 8,380,540 light chain and 4,925,532 heavy chain sequences (referred to as "Ig-seq set" in this chapter). We carried out the SCALOP and FREAD prediction and analysis on all of the redundant sequences. For unique CDR loops, there are 513,143 H1, 523,536 H2, 353,229 L1, 77,689 L2 and 496,229 L3. The "overlap coverage" is the percentage of sequences for which both FREAD and SCALOP made a prediction. Within the overlapped predictions, if the FREAD prediction was <1.5Å backbone RMSD from any member of the cluster assigned by SCALOP, it was considered a "consistent prediction".

2.2.6 Backdating the SCALOP database

We used the whole SAbDab set as the test set. For the back-dated set, we retained structures whose deposition dates were before the end of the year of interest. We selected the representative years based on previous publication dates of canonical forms definitions (Al-Lazikani et al., 1997; North et al., 2011; Nowak et al., 2016). For each back-dated set, we carried out a leave-one-out cross-validation for all the loops in the SAbDab set:

- for loops that existed on or before the given year, we did not include the loop of interest in the construction of PSSMs, and
- for loops that came into existence later, we built the PSSMs based on all loops present in the back-dated set.

2.2.7 Observing the correlation between CDR canonical forms and the antibody binding

We tested for a correlation between the CDR canonical classes predicted by SCALOP and their relationship with the binding affinity. Adams et al. (2016) curated a dataset of mutated H1 sequences and the resultant antibodies' binding affinity to the

target antigen. In the experiment, single, double or triple mutations were introduced to the H1 sequence of a fluorescein-binding antibody (PDB code: 1FLR; referred to as wild-type/WT). They used a high throughput binding affinity measurement technique, Tite-seq, to calculate the affinity (K_D) of these ∼2800 antibodies with mutated H1 sequences. A variant with femtomolar affinity (PDB code: 1X9Q; referred to as OPT) was also assayed.

The dataset was curated using the CDR definition according to Kabat and Wu (1991). Therefore we constructed the PSSMs of SCALOP according to this definition. SCALOP was then used to predict the canonical classes of all these mutated CDRH1 sequences. The measured binding affinity and the predicted canonical classes were then plotted to see if the changes in the canonical class influenced the binding affinity.

2.3 Results and Discussion

2.3.1 Performance of SCALOP and FREAD on SAbDab set

We evaluated the performance of SCALOP on the SAbDab set using a leave-one-out cross-validation protocol and compared its performance to FREAD. We compared with FREAD because it is used as a high quality, high coverage CDR structural predictor (Leem et al., 2016). The prediction coverage and precision of both methods for all non-H3 CDR loops are shown in Table 2.2. In general, SCALOP has slightly poorer coverage than FREAD. The average coverage for SCALOP across all CDRs is 95.8% against 97.2% for FREAD. This is likely to be caused by a lower coverage of the structures captured by the SCALOP clusters, as some loops fall outside of any canonical cluster, preventing us from predicting structures of that type. On the other hand, FREAD does not depend on cluster formation but surveys the full structural database. The precision of FREAD and SCALOP are both high (average 90.19% and 94.19% respectively), suggesting that SCALOP predictions can be used

to rapidly predict the structures of non-H3 CDR loops.

Table 2.2: Coverage and precision of SCALOP and FREAD on SAbDab set.

Metric	Method	H1	H2	L1	L2	L3
Coverage	SCALOP	93.75	97.54	97.38	98.50	91.69
	FREAD	96.79	93.38	98.76	98.89	98.02
Precision	SCALOP	89.26	93.60	95.67	99.13	93.31
	FREAD	80.19	88.50	92.72	98.27	91.29

2.3.2 Performance of SCALOP and FREAD on the Ig-seq set

We ran SCALOP and FREAD on the Ig-seq set and assess their overlap coverage, consistency and speed. Figures 2.2 and 2.3 present the prediction coverage of SCALOP and FREAD on the Ig-seq data. Between 78.58% and 98.02% of the predictions for the five CDRs overlap between SCALOP and FREAD. FREAD once again gives higher coverage than SCALOP. In heavy chain CDRs (Figure 2.2), FREAD has approximately 20% (H1) and 16% (H2) more coverage than SCALOP, compared to the 2% (L1) and 1% (L2) in light chain counterparts (Figure 2.3A–B). There is also a small portion of loops (<1%) which were not predicted by FREAD but were predicted by SCALOP. In most of the CDRs (Figures 2.2 and 2.3A–B), apart from L3, only a tiny portion of the loops were predicted by neither FREAD nor SCALOP (0.11% to 0.58%). In L3 loops (Figure 2.3C), FREAD has approximately 10% more coverage than SCALOP. No FREAD or SCALOP predictions were made for 1.75% of the loops, a percentage that is considerably higher than the <1% in the other CDR types. We had expected L3 loops to be slightly more diverse, thus comparatively less predictable, than the other four CDRs, owing to the junctional diversity added to the third hypervariable loop of each chain as described in Section 1.2.1.1.

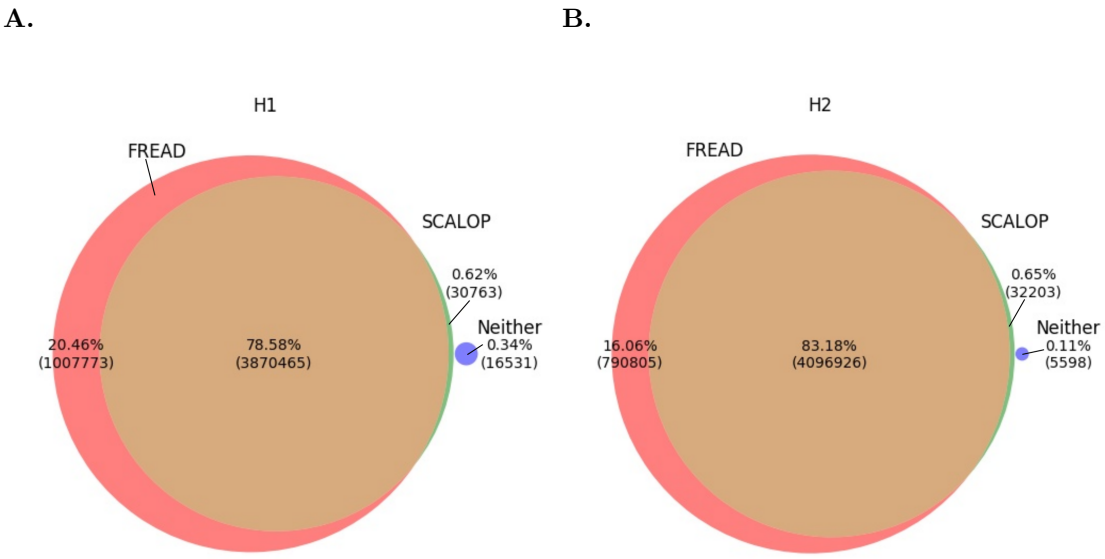
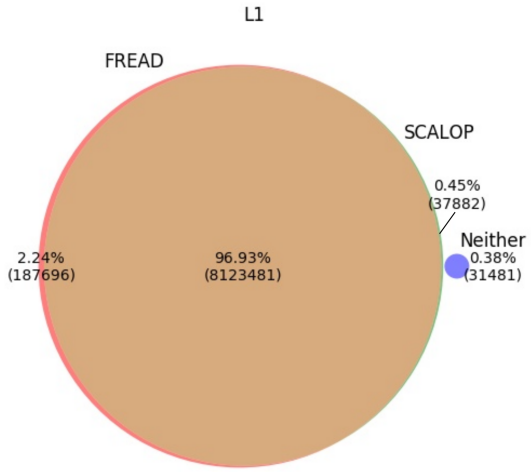


Figure 2.2: Prediction coverage of SCALOP and FREAD on the Ig-seq data, for each heavy chain CDR types. FREAD has a higher coverage than SCALOP in all cases. **A.** CDRH1. The overlap coverage is close to 80%. Only 0.34% of sequences were not predicted by either FREAD nor SCALOP. **B.** CDRH2. The overlap coverage is above 80%. Only 0.11% of all sequence data is predicted by neither method.

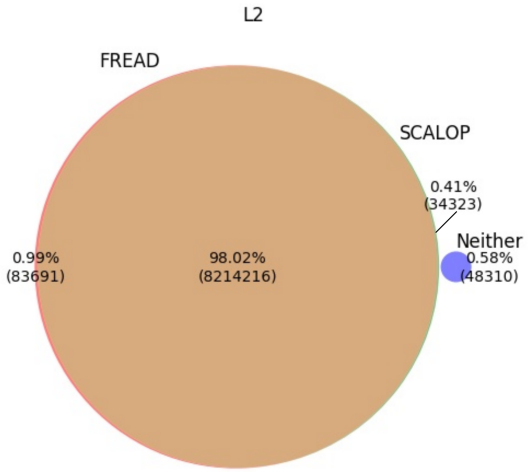
The percentage of consistent predictions affirms that SCALOP makes very similar predictions to FREAD (Table 2.3). For CDRs on the light chain, all predictions made by SCALOP and FREAD agree. In heavy chain CDRs, the 5% disagreement could be caused by the fewer number of clusters in the heavy chain CDRs leading to less informative PSSMs. It is also possible that, since SCALOP and FREAD are predictors, either, or both of the methods could make incorrect predictions in this comparison.

One of the major advantages of SCALOP is its speed. On a single core, predicting 100 sequences takes 227s using FREAD, but 0.292s using SCALOP. This nearly 800-fold speed up suggests the possibility of running SCALOP as a fast and reliable first-screen for a large sequencing dataset.

A.



B.



C.

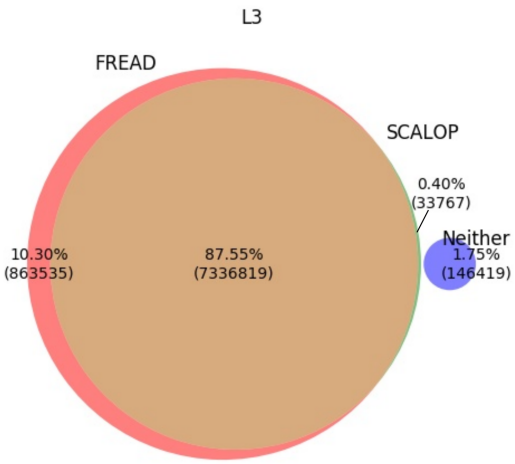


Figure 2.3: Prediction coverage of SCALOP and FREAD on the Ig-seq data, for each light chain CDR types. FREAD has a higher coverage than SCALOP in all cases. **A.** CDRL1. The coverage of SCALOP and FREAD are comparable while the overlapping coverage is close to 97%. **B.** CDRL2. SCALOP and FREAD make predictions over almost the same set of loops. **C.** CDRL3. The overlap coverage is reasonably high for a loop that is marginally more variable than the other CDRs with canonical forms.

Table 2.3: Overlap coverage and consistent prediction within the overlap in Ig-seq set.

CDR	Overlap Coverage (%)	Consistent prediction (% of overlap coverage)
H1	78.58	95.15
H2	83.18	95.24
L1	96.93	100
L2	98.02	100
L3	87.55	100

2.3.3 Backdating the SCALOP database: Performance evaluation

In order to ensure that SCALOP always offers the best possible prediction coverage, it uses an auto-updating database. To illustrate the importance of maintaining an auto-updating database, we back-dated the SCALOP database and assessed the coverage and precision. The results (Figures 2.4 and 2.5) show that by updating the database, prediction coverage increases while retaining high precision.

For CDRH1 (Figure 2.4A), we observed only one canonical form in 1997. As the number of sequences increased by 20-fold within 20 years, we have found three clusters in 2016. Length-14 and length-15 clusters were absent in 1997, but appeared from 2011. The prediction coverage increased slightly from nearly 80% to above 90%, and the precision maintained at 89%.

A near 40% increase of CDRH2 canonical form prediction coverage and an improvement of ∼10% in precision are seen between 1997 and 2016 (Figure 2.4B). The portion of non-clustered sequences (16.8%) has dropped to one-third of the portion in 1997 (55.2%). Both clusters in 1997 were from 10-residues long loops. In 2016, length-9 and length-12 loops also had their own canonical forms, arising from a similar 20-fold expansion in the number of CDRH2 sequences as observed in CDRH1.

We found the most number of clusters in CDRL1. The number of clusters grew from four in 1997 to twelve in 2016 (Figure 2.5A). Length-17 sequences joined the 1997's length-16 cluster to form the L1-16,17-A cluster. The 2011's L1-11-C cluster combined with the 2011's L1-11-B cluster and resulted in the 2016's L1-11-B cluster. The portion of non-clustered sequences in 2016 (7.7%) dropped to a quarter of the portion in 1997 (47.8%). Prediction coverage increased by over 40% between 1997 and 2016, from 54.7% to 97.4%, while the precision retained at ∼95%.

We observed that the conformation of CDRL2 loop is largely invariant – most loops only belong to a single length-8 cluster (Figure 2.5B). Some length-12 CDRL2 loops appeared in the later years but were not clustered. This minimal perturbation in the structural understanding of CDRL2 is evident in the consistently high prediction coverage and precision.

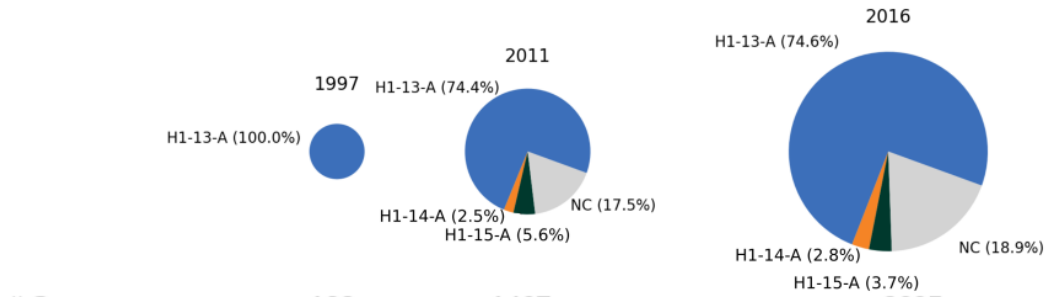
In CDRL3 loops, only one length-9 cluster existed in 1997, compared to seven in 2016 (Figure 2.5C). Between 2011 and 2016, some length-10 sequences joined the 2011's L3-9-A cluster, which becomes the 2016's L3-9,10-A cluster. Likewise, some length-10 loops joined the 2011's L3-11-A cluster to form 2016's L3-10,11-A cluster. A near 30% increase in prediction coverage is observed with good precision above 90%.

Overall, as more structures are included, some clusters increase in size, new clusters form from previously non-clustered loops and some clusters combine to form a single cluster. This shows how the increase in structural data enriches our knowledge in CDR canonical forms. The database of a canonical form predictor needs to be updated continuously to capture these changes in our understanding of canonical forms.

Despite having different training sets for each year, the precision of SCALOP prediction remains around 90%. The robustness of the SCALOP to predict canonical

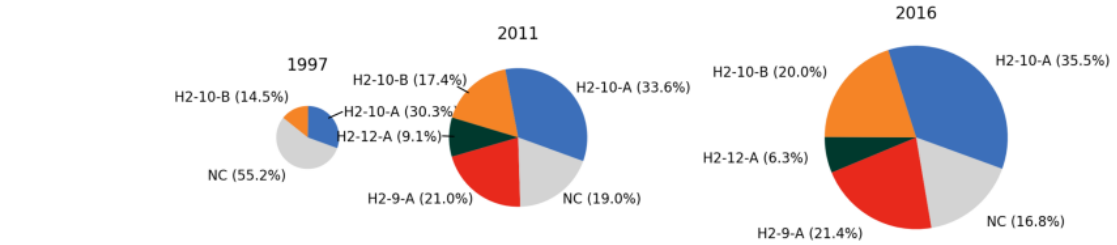
forms is confirmed by the independence of the precision from the training set.

A.



	1997	2011	2016
# Sequences	123	1467	2695
Coverage (%)	79.1	93.2	93.8
Precision (%)	88.9	88.6	89.4

B.



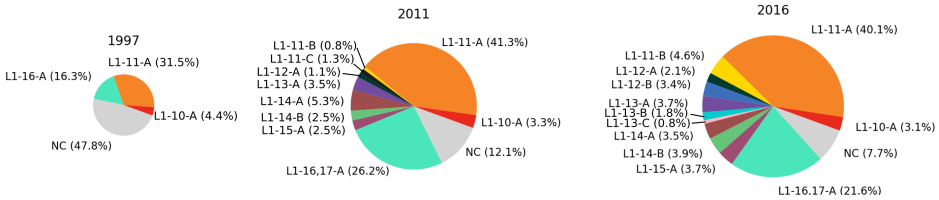
	1997	2011	2016
# Sequences	145	1487	2768
Coverage (%)	58.3	95.9	97.5
Precision (%)	82.5	91.9	93.6

Figure 2.4: The changes in heavy chain CDRs cluster composition from 1997 to 2016 and their prediction coverage and precision. The radii of the pie charts are proportional to the log(#Sequences). NC refers to non-clustered sequences. **A.** CDRH1. The number of canonical forms increase from one to three, with a 20-fold increase in the number of sequences. **B.** CDRH2. A large increase in prediction coverage is seen over 20 years, with a drop in the portion of NC sequences.

2.3.4 Changes in CDR canonical forms from the native structure is likely to impact the antibody binding affinity

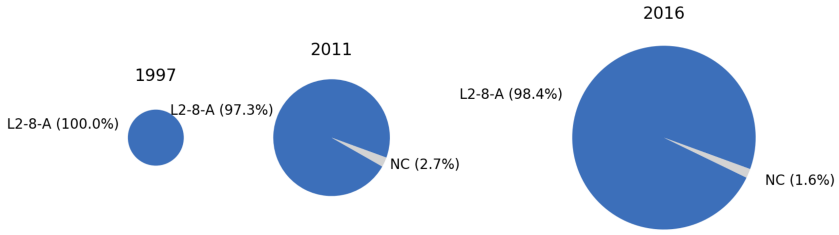
As described in Section 2.2.7, we examined the canonical forms of a mutagenesis and binding affinity dataset curated by Adams et al. (2016). We made canonical form predictions on ∼2800 mutated CDRH1 sequences (in Kabat definition; Kabat and

A.



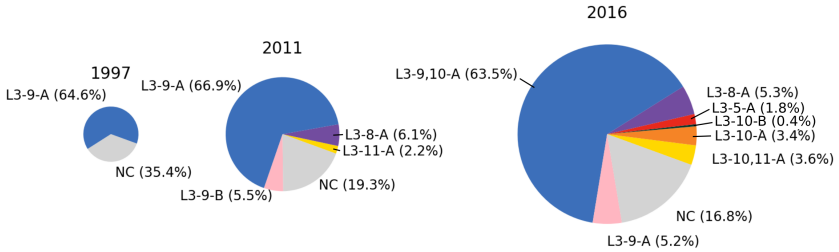
	1997	2011	2016
# Sequences	203	1500	2567
Coverage (%)	54.7	93.0	97.4
Precision (%)	96.5	94.8	95.7

B.



	1997	2011	2016
# Sequences	199	1571	2724
Coverage (%)	98.5	98.5	98.5
Precision (%)	98.9	99.1	99.3

C.



	1997	2011	2016
# Sequences	206	1532	2672
Coverage (%)	62.5	79.0	91.6
Precision (%)	96.3	95.3	93.4

Figure 2.5: The changes in light chain CDRs cluster composition from 1997 to 2016 and their prediction coverage and precision. The radii of the pie charts are proportional to the log(#Sequences). NC refers to non-clustered sequences. **A.** CDRL1. The number of clusters grew from 4 in 1997 to 12 in 2016. Length-17 sequences merged with length-16 cluster in 1997 to form the L1-16,17-A cluster. The portion of non-clustered sequences dropped to a quarter of the portion in 1997. **B.** CDRL2. The conformation of L2 loop is largely invariant, hence most loops only belong to a single cluster. **C.** CDRL3. In 1997, only one cluster existed. An increase in prediction coverage is observed with good precision above 90%.

Wu, 1991) and observed how the changes in canonical forms affected the binding affinity. Figure 2.6 shows the measured binding affinity (K_D) and canonical forms predicted by SCALOP. The CDRH1 sequences in wild-type (WT) and "optimal" (OPT) binders belong to the same canonical form (H1-5-A). Where the mutations caused a change from this canonical form, the binding tended to become less favourable, except in a single circumstance where the sequence was predicted to adopt the shape in the H1-5-C cluster. However, within H1-5-A canonical class, both strong and weak binders are found. The case study shows that SCALOP can highlight structural changes that can impact binding.

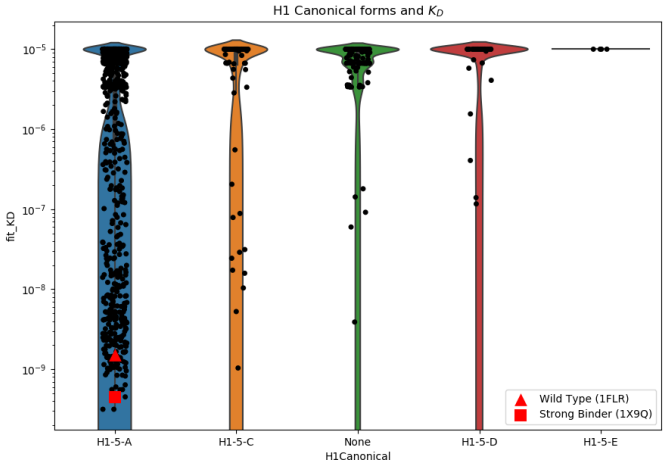


Figure 2.6: H1 canonical forms of the mutated sequences predicted by SCALOP and the K_D measured by Adams et al. (2016). The red triangle and square indicate the data points for the WT (1FLR) and OPT (1X9Q) respectively. Each black point represent a mutated sequence and its measured binding affinity. The violin plot for each cluster indicates the data distribution within the cluster.

2.4 Chapter Summary

In this chapter we have described SCALOP, a sequence-based antibody CDR canonical form annotation tool. We evaluated its prediction performance and benchmarked with a database-search loop prediction tool, FREAD (Choi and Deane, 2010), to show that SCALOP and FREAD give consistent predictions on

an Ig-seq dataset. By back-dating the training set of SCALOP, we showed that the PSSM algorithm is robust to the ever-changing training set. We further confirmed an increasing number of CDR canonical forms over the years, and concluded that an automatic update to SCALOP's training database is necessary. Finally we applied SCALOP to annotate a mutagenesis dataset (Adams et al., 2016) and observed a deviation of canonical forms could impact binding affinity.

SCALOP has the advantage of scalability on large sequencing datasets and it gives consistent structural annotations with FREAD. There are two possible ways to combine the prediction power of SCALOP and FREAD. SCALOP could act as a fast first screen on a large sequence dataset to provide a structural approximation to the majority of the sequences, and FREAD can be used to predict loops which fall outside of SCALOP's prediction. This minimises the portion of data which requires the slower FREAD prediction. This pipeline has been implemented in Kovaltsuk et al. (2020), to show that by replacing the FREAD prediction step on the five canonical CDRs with SCALOP (Krawczyk et al., 2018), it is possible to annotate 4.5 million antibody sequences in a day on a 40-core computer.

Furthermore, if a structural model is required, SCALOP could be used to make a cluster prediction which could act as a constraint for the FREAD database-search algorithm. This has been implemented as an additional feature to better visualise the SCALOP prediction results on the web server.

In the next chapter, we will transfer the SCALOP framework on analyse T-cell receptor (TCR) CDRs, and compare the structural differences in antibody and TCR CDRs. The observation inspires the development of a multi-state TCR homology modelling server, TCRBuilder, that is based on the ABodyBuilder pipeline (Leem et al., 2016) and captures the structural variability in TCR CDRs.

3

Comparative analysis of the CDR loops of antigen receptors and multi-state TCR homology modelling

Contents

This chapter has been adapted from my published work "Comparative analysis of the CDR loops of antigen receptors" in *Frontiers in Immunology* (Wong et al., 2019b) and "TCRBuilder: multi-state T-cell receptor structure prediction" in *Bioinformatics* (Wong et al., 2020a).

3.1 Introduction

Antibody and T-cell receptor (TCR) have similar genetic mechanism and architecture, as described in Section 1.2.3. In both types of antigen receptors, sequence

and structural diversity is concentrated in six complementarity determining regions (CDRs). There are three in the TCRα chain (CDRα1–CDRα3) and three in the TCRβ chain (CDRβ1–CDRβ3). Likewise, the light chain and heavy chain of antibodies have three CDRs each (CDRL1–CDRL3, CDRH1–CDRH3). These CDRs form the majority of the binding site.

3.1.1 TCR CDR canonical forms

As described in Section 1.2.1.3, the canonical model for antibody CDRs has proven a powerful tool in antibody design and modelling. Despite this, only two studies have so far applied the concept to TCR CDRs (Al-Lazikani et al., 2000; Klausen et al., 2015).

The first clustering of TCR CDR loops was carried out using only seven TCR structures (Al-Lazikani et al., 2000). At that time, four canonical classes were identified for CDRα1, four for CDRα2, three for CDRβ1, and three for CDRβ2; neither CDR3 loop was clustered. More recently, Klausen et al. (2015) clustered the CDRs from a non-redundant set of 105 $\alpha\beta$TCR and 11 unpaired TCR structures. They performed the clustering in torsion space using an affinity propagation algorithm. In total, 38 canonical forms were characterised. These clusters were then used to construct a sequence-based, random forest classifier, with canonical form prediction accuracies between 63.21% and 98.25% (Klausen et al., 2015).

3.1.2 Anchor residue analysis in CDRβ3/CDRH3

CDRβ3 in TCRs and CDRH3 in antibodies show higher variability in sequence composition and structure than the other CDRs (Janeway et al., 2001). While no canonical forms have been defined for CDRH3, several groups have analysed the kinked and extended (or bulged and non-bulged) conformations at the start and end of the loop, known as the “base” or “torso” region (Shirai et al., 1996; Morea

et al., 1998; Shirai et al., 1999; Kuroda et al., 2008; North et al., 2011; Weitzner et al., 2015; Finn et al., 2016). Weitzner et al. (2015) showed that pseudo bond angle τ and pseudo dihedral angle α of the second last residue of the CDRH3 loop (Chothia position 101, IMGT position 116; Al-Lazikani et al., 1997; Lefranc et al., 2015) can differentiate between the extended and kinked torso conformations. Finn et al. (2016) analysed the first three and last four residues of CDRH3 loops and observed that for the same IMGT position 116, the ϕ/ψ angles are different in kinked and extended torsos. In this chapter, we carry out the first examination of the conformation of the base region of CDRβ3 loops.

3.1.3 Comparative studies between TCR and antibody CDRs

Although TCRs and antibodies are derived from similar genetic mechanisms and share a similar architecture, only a handful of studies have compared them (*e.g.* Rock et al., 1994; Bentley et al., 1995; Dunbar et al., 2014a; Allison et al., 2001; Lefranc et al., 2015; Blevins et al., 2016). Furthermore, analyses have largely focussed on sequence-based features. For instance, Rock et al. (1994) found that the CDRα3 and CDRβ3 loops have a different length distribution to CDRL3 and CDRH3, while Blevins et al. (2016) observed that the TCR CDR1 and CDR2 sequences have more charged amino acids than the analogous antibody CDR1 and CDR2. Given the similarity in the genetic mechanisms, fold and the limited conformational variability in the canonical CDRs, it might be reasonable to expect structural similarity between TCRs and antibodies. Comparing TCRs and antibodies should identify potential characteristics that may inspire antibody-like TCR design, TCR-mimic antibodies and soluble TCRs (Dunbar et al., 2014a; Dubrovsky et al., 2016; Trenevska et al., 2017; Xu et al., 2018). Such analyses can also highlight structural signatures that may relate to their different biological functions, such as major histocompatibility complex (MHC) restriction in TCRs (Attaf et al., 2015) and the virtually unconstrained antigen binding in antibodies (Sela-Culang et al., 2013).

3.1.4 TCR structural modelling

Structural modelling may provide further insights into the binding of TCRs. Several packages exist to model the structures of TCRs including LYRA, TCRModel and Repertoire Builder (Klausen et al., 2015; Gowthaman and Pierce, 2018; Schritt et al., 2019). All of these methods use homology modelling techniques (see Section 1.2.2.1) and they only take a single snapshot of the predicted structure. A number of experimental studies (Garcia et al., 1998; Morris and Allen, 2012) and our analysis of TCR CDRs in Section 3.3.4 suggest that TCR polyspecificity could be linked to the highly variable CDR structures. The ability to capture the multiple conformations of CDRs may be necessary for TCR structural modelling to better understand its binding mechanism.

3.1.5 Chapter Overview

In this chapter, we analyse the CDR structures in antibody and TCRs, and describe a novel TCR structural modelling tool. We first used a length-independent clustering method to update the canonical classes of TCR CDRs (Nowak et al., 2016). We then built a sequence-based TCR CDR canonical form prediction algorithm based on an adapted position-specific scoring matrix (Wong et al., 2019a) outlined in Chapter 2. Next, we attempted to enrich the TCR CDR dataset with antibody CDR structures and found that TCR and antibody CDRs occupy distinct areas of structural space. In the small number of common conformational clusters, we found that the underlying sequence patterns differ. Our examination also showed that the structural variability in TCR CDRs is higher than that in antibody CDRs. To capture the multiple conformations in TCR CDRs, we developed TCRBuilder, a multi-state TCR modelling server. TCRBuilder is based on the ABodyBuilder pipeline for antibody modelling (Leem et al., 2016) and predicts TCR models with multiple CDR conformations where applicable. The canonical form prediction tool (SCALOP-TCR) and TCRBuilder are available at

`http://opig.stats.ox.ac.uk/webapps/stcrpred`.

3.2 Method

3.2.1 Nomenclature

We use the following nomenclature for structures: four characters of the Protein Data Bank (PDB; Berman et al., 2000) code, followed by a "_", then the chain identifier of the structure, *e.g.* 5HHM_E.

CDR clusters are identified by two letters describing the CDR type, the loop length(s), then a letter (Nowak et al., 2016). For instance, the α1-6-B class refers to the length-6 CDRα1 class with the second-largest number of unique sequences. Pseudo-clusters (see Section 3.2.4) with only one unique sequence have their names appended with an asterisk (*). For the clustering of both TCR and antibody CDRs, we name the clusters by the two letters representing the TCR CDR type, two letters for the antibody CDR type, the loop length(s), followed by a letter, *e.g.* α1,L1-11,12-A denotes the largest cluster from the CDRα1/CDRL1 clustering, in which sequences are 11 or 12 residues long.

3.2.2 Definitions

TCR and antibody structures were numbered in the IMGT scheme (Lefranc et al., 2015) using ANARCI (Dunbar and Deane, 2016): CDR1 (27–38), CDR2 (56–65), and CDR3 (105–117). To compare the CDR loops of TCRs and antibodies, we assumed equivalence between β/heavy chains and α/light chains (Leem et al., 2018a), as TCRβ and antibody heavy chains rely on VDJ recombination, while TCRα and antibody light chains are formed by VJ recombination.

3.2.3 Structural datasets for CDR analysis

TCR structures with at least one α or β chain and resolution $\leq$2.8Å were downloaded from STCRDab (Leem et al., 2018a) on 31$^{\text{st}}$ May 2018. Antibody structures with

resolution ≤2.8Å were downloaded from SAbDab (Dunbar et al., 2014b) on 31st May 2018. Structures from all species were retained. In total, 270 TCR PDB entries and 2563 antibody PDB entries were used. We retained all chains in a PDB entry that passed the quality criteria. The structures of the CDR loops were extracted using a similar procedure to Nowak et al. (2016). The IMGT-defined CDRs (Lefranc et al., 2015), along with five N-terminal and five C-terminal anchor residues, were selected if there were no missing backbone atoms and none of the backbone atoms had B-factors >80. Loops also needed to be continuous, *i.e.* all peptide bond lengths <1.37Å. Since loops with identical sequences can adopt multiple conformations (see Section 3.3 and Figure 3.3 in the current work on TCR CDRs and examples in Nowak et al., 2016, for antibody CDRs), we retained all CDR loop structures that passed the structural quality criteria. This set of TCR structures is referred to as "STCRDab 2018 set", and the antibody structures the "SAbDab 2018 set".

3.2.4 Structural clustering method

Loops are clustered using the length-independent clustering method from Nowak et al. (2016), as described in Section 2.2.1.2.

In the clustering of TCR CDRs alone, we only consider a set of sequences to be a CDR cluster if it contains a minimum of five structures and two unique sequences. If a set contains five or more structures but only one unique sequence, we label this a "pseudo-class". All other loops are considered to be "non-clustered". This is a more lenient threshold than previous investigations clustering antibody CDR loops (*e.g.* Nowak et al., 2016), as the dataset of TCR structures is far smaller. To choose the optimal clustering parameter, DBSCAN was run over a range of DTW score thresholds in increments of 0.1Å. We find that 1.0Å offered an optimal balance between the number and size of clusters for the five TCR CDR loops.

For the clustering of multiple CDR types, we use the same clustering thresholds and criteria for the respective CDR types as the antibody study (Nowak et al.,

2016), where they were selected using the Ordering Points to Identify the Clustering Structure (OPTICS; Ankerst et al., 1999) algorithm. CDRα1/CDRL1 are clustered at 0.82Å, CDRα2/CDRL2 at 1Å (not initialised in the previous paper), CDRα3/CDRL3 at 0.91Å, CDRβ1/CDRH1 at 0.8Å and CDRβ2/CDRH2 at 0.63Å. A valid cluster must contain at least six unique sequences as illustrated in Nowak et al. (2016).

3.2.5 Comparison with TCR canonical classes in earlier work

In order to compare our CDR clusters to previous studies (Al-Lazikani et al., 2000; Klausen et al., 2015), we identified overlaps in the representative PDB entries. For example, Al-Lazikani et al.'s α2-3 class contains the PDB entry 1TCR (Al-Lazikani et al., 2000). Since 1TCR_A was in our α2-8-B class, we considered these two classes (α2-3 and α2-8-B) to be analogous. A similar procedure was used to map our classes to Klausen et al.'s classes (Klausen et al., 2015).

Some canonical classes from Al-Lazikani et al. (2000) or Klausen et al. (2015) were represented by structures that were filtered out of our dataset. To match these classes, we searched for a CDR structure in our set that has the same CDR sequence and checked if there was a backbone match. For example, the centroid structure of Klausen et al.'s α3-7 class is PDB 3TF7, which has the sequence AVSAKGTGSKLS. We found that 3TFK_C has the same CDRα3 sequence as 3TF7 and a backbone root-mean-square deviation (RMSD) of 0.81Å; thus, we assigned their α3-7 to our α3-12-A class.

3.2.6 Sequence-based prediction of canonical forms

Similar to the method described in Section 2.2.2, we built a sequence-based predictor for the TCR structural clusters. Hereafter, the TCR version of SCALOP is called "SCALOP-TCR".

3.2.6.1 Building the position-specific scoring matrix

For each CDR canonical class, we generated a position-specific scoring matrix (PSSM). The score of an amino acid k in IMGT position j, $M_{k,j}$, is

$$M_{k,j} = \log_2 \frac{p_{k,j}}{b_k}, \tag{3.1}$$

where $p_{k,j}$ represents the probability of k at j, and this is calculated separately for each class. The background probability of k, b_k, was assumed to be identical for all residues *i.e.* 0.05.

3.2.6.2 Scoring the PSSM and making a prediction

To predict the cluster for a target loop with length l, we first select PSSMs containing loops with the same length. For example, if the new CDRα1 loop is six residues long, we choose PSSMs of the α1-6-A, α1-6-B, α1-6-C and α1-6-D classes. The PSSM score for the target loop for class c, s_c, is the sum of the position-specific scores:

$$s_c = \sum_{j=J_0}^{J} M_{k,j}. \tag{3.2}$$

If k is never observed at j, we assume $M_{k,j} = -1$. For canonical class assignment, we designated a target loop to the class with the highest value of s_c. Furthermore s_c must be higher than 1, except for CDRα3 where s_c must be equal to or greater than the loop's length. The value of s_c was chosen by performing leave-one-out cross-validation tests over several values of s_c. Sequences were assigned to a pseudo-class if and only if they had an identical sequence.

3.2.6.3 Cross-validation

To benchmark the scoring strategy, we ran a leave-one-out cross-validation protocol on the unique sequences from canonical classes and non-clustered structures. A prediction was evaluated using the following criteria:

- True positive: sequence is assigned to the correct canonical class.

- False positive: sequence is assigned to a different canonical class.
- True negative: sequence is from a non-clustered loop, and not assigned a canonical class.
- False negative: sequence is from a canonical class, but predicted to be non-clustered.

We evaluated the precision and recall using Equations 2.4 and 2.5.

3.2.6.4 Prediction on Ig-seq dataset

We predicted the canonical forms for the α-chain in a set of mouse TCR sequences from BioProject accession PRJNA362309 (Maceiras et al., 2017). Overlapping Illumina reads were assembled using FLASh (Magoč and Salzberg, 2011), and TCR amino acid sequences were extracted using IgBLAST (Ye et al., 2013). The sequences were then numbered by ANARCI (Dunbar and Deane, 2016); only those with productive CDR3 rearrangement, CDR1 and CDR2 loops at least five residues long, and CDR3 loops at least eight residues long were retained.

3.2.6.5 Prediction of new TCR structures

We used the 44 $\alpha\beta$TCR structures that were released between 31$^{\text{st}}$ May 2018 and 5$^{\text{th}}$ June 2019 as a blind test set ("SCALOP-TCR blind set"; see Table 3.1). Unlike our structural dataset for clustering, we did not impose a quality restriction for CDR prediction. Predictions were considered to be correct if the backbone RMSD between the native CDR structure and any member of the assigned canonical class was $\leq$1.0Å.

3.2.7 Comparison between TCR and antibody CDR structures

We clustered TCR and antibody CDR structures and examined the length distribution, structural clustering and sequence patterns. Sequence lengths of CDR loops disregard the five anchor residues at each side of the CDR structures. Structural clustering comparison were done as described above. We used WebLogo to generate

Table 3.1: List of PDB structures used in SCALOP-TCR blind set, and any structures that share identical sequences where applicable. PDB codes are listed as PDB ID followed by the TCR chain IDs.

PDB Codes in the blind set			
6MIV_A	6MIY_H	6MJA_C	6MJQ_B
6MIV_B	6MJ4_A	6MJA_D	6MJQ_C
6MIV_C	6MJ4_B	6MJI_A	6MJQ_D
6MIV_D	6MJ4_C	6MJI_B	6MJQ_E
6MIY_A	6MJ4_D	6MJI_C	6MJQ_F
6MIY_B	6MJ6_A	6MJI_D	6MJQ_G
6MIY_C	6MJ6_B	6MJJ_A	6MJQ_H
6MIY_D	6MJ6_C	6MJJ_B	6MTM_A
6MIY_E	6MJ6_D	6MJJ_C	6MTM_B
6MIY_F	6MJA_A	6MJJ_D	6MTM_D
6MIY_G	6MJA_B	6MJQ_A	6MTM_E

all sequence patterns (Crooks et al., 2004). Amino acids were coloured with the default hydrophobicity scale: hydrophilic residues (R, K, D, E, N and Q) were in blue, neutral residues (S, G, H, T, A and P) in green and hydrophobic residues (Y, V, M, C, L, F, I and W) in black.

In canonical forms where both TCR and antibody CDRs were found, we examined the difference between the sequence patterns used by TCRs and antibodies. We transformed the sequences using one hot encoding by position, where each feature was represented by the position and the residue name, and applied Principal Component Analysis (PCA) using the *scikit-learn* module in Python (Pedregosa et al., 2011).

3.2.8 Analysis of CDRβ3 and CDRH3 structures

We analysed CDRβ3 structures with a set of metrics that have been previously applied to the CDRH3 loop in antibodies: loop anchor transform (LAT), pseudo bond angles and dihedral angles (Weitzner et al., 2015; Finn et al., 2016).

LAT is the Euler transformation of the coordinate planes formed by residues at IMGT positions of 105 and 117 (see Supplementary Information of Weitzner et al., 2015, for a detailed mathematical definition). Briefly, a coordinate system is defined for each of the two residues centred on the C_α atoms, where the z-axis points towards the carbonyl carbon, the y-axis is perpendicular to z in the N-C_α-C plane, and the x-axis is the cross product of these two components. The Euler transformation is then represented by six degrees of freedom, capturing the translation (X, Y, Z) and rotation (ϕ, ψ, θ).

We calculated the pseudo bond angle (τ) and pseudo dihedral angle (α) of the residue at IMGT position 116 (corresponding to Chothia position 101), in the CDRβ3 and CDRH3 sets. Consistent with Weitzner et al. (2015), the pseudo bond angle is formed by the C_α atoms of the residues before, at and after the IMGT position 116, whereas the pseudo dihedral angle spans across the C_α atoms of residues at position 116, one before and two after.

Finn et al. (2016) observed that the dihedral angles adopted by the base residues of extended torsos were different from that of kinked torsos. To capture the observation made by Finn et al. (2016), the dihedral angles of the first three (T1-T3) and last four (T4-T7) residues of the CDRβ3 and CDRH3 loops are obtained from the Biopython module (Cock et al., 2009).

3.2.9 TCRBuilder: Multi-state TCR structure prediction

3.2.9.1 TCRBuilder Database

A database of $\alpha\beta$ T-cell receptor (TCR) structures that were solved by X-ray crystallography, at resolution $\leq$ 3.2Å, were extracted from STCRDab (Leem et al., 2018a), on 23$^{\text{rd}}$ July 2019. We refer to this as the "STCRDab 2019 set".

When building the CDR type-specific FREAD databases (Deane and Blundell, 2001; Choi and Deane, 2010; see Section 2.2.4), IMGT-CDR structures (Lefranc et al., 2015) were extracted with five residues before and five after the CDR. Fragment lengths range between 3 and 30.

3.2.9.2 Modelling pipeline

The modelling pipeline of TCRBuilder is based on the ABodyBuilder pipeline (Figure 3.1; Leem et al., 2016).

The $\alpha\beta$TCR target sequences are numbered using ANARCI (Dunbar and Deane, 2016). For each individual query chain, a framework template is selected from STCRDab (Leem et al., 2018a) using the highest sequence identity over the framework region. The orientation template is then chosen from the structure with the highest full sequence identity (across both the framework and the CDRs) over both chains in the pair. Any missing residues in the framework are modelled using MODELLER (Webb and Sali, 2014).

CDRs are modelled in the following order: $\alpha1$, $\beta1$, $\alpha2$, $\beta2$, $\alpha3$ and finally $\beta3$. FREAD (Deane and Blundell, 2001; Choi and Deane, 2010), a database-based loop prediction method, searches through its database of the corresponding CDR type and returns a list of hits. If the exact query CDR sequence, or the sequence of the top hit (ranked by the anchor RMSD) is found in multiple distinct conformations, all unique conformations are retained and considered as alternative models. That is, if two structurally distinct templates were found for $\alpha1$, each of the structures will proceed through an independent modelling pipeline of the $\beta1$, $\alpha2$ and so on. If there is only one conformation with an exact sequence match, or the top hit if there is no exact match, one model is selected to proceed with the modelling. Here distinct conformations refer to unique conformations whose backbone atom (C, C_α, N and O) RMSDs differ by >1Å. Where FREAD makes

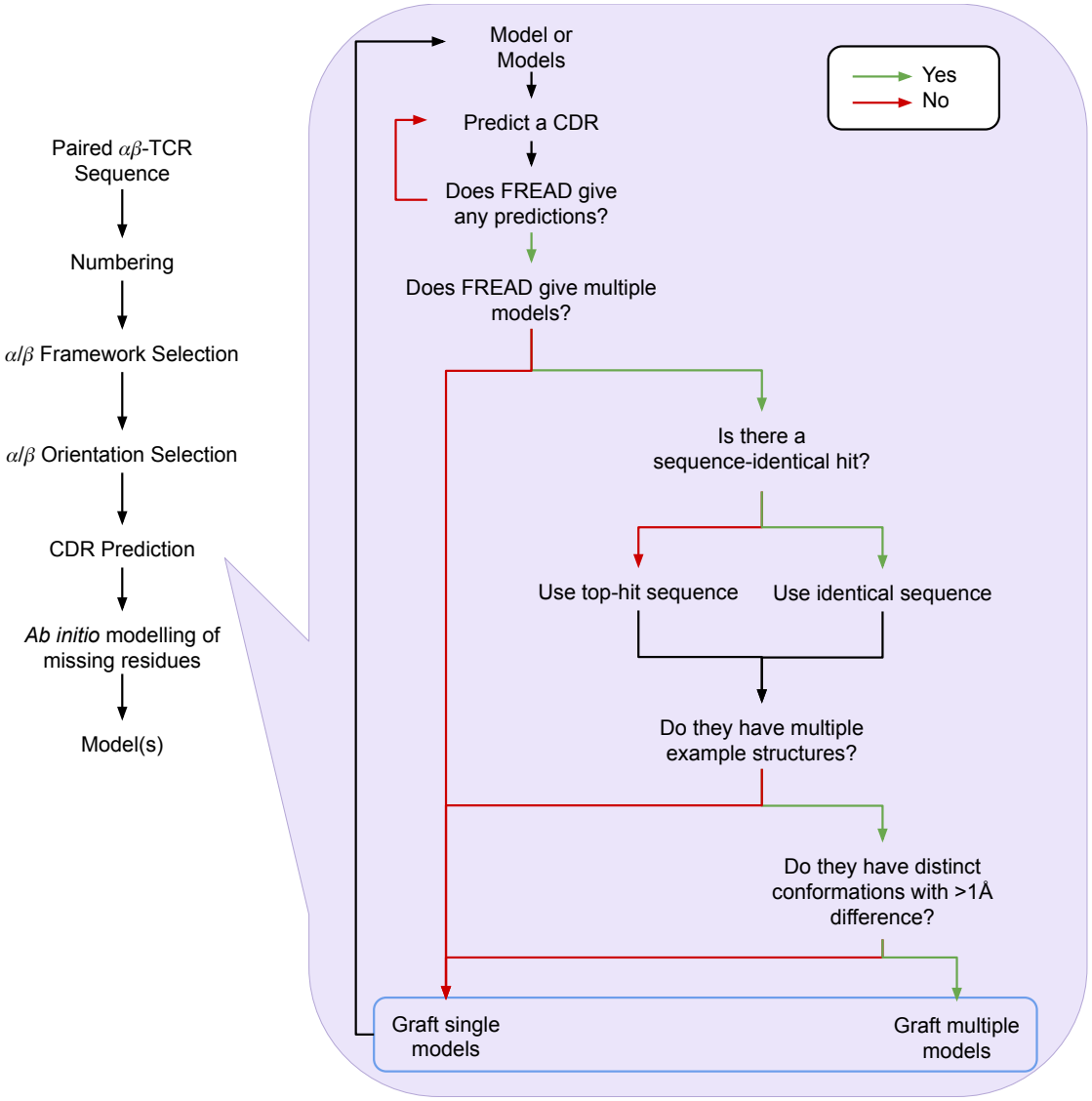


Figure 3.1: TCRBuilder pipeline for modelling TCR from sequence. The target $\alpha\beta$TCR sequences are numbered by ANARCI (Dunbar and Deane, 2016). Framework and orientation templates are selected by sequence identity. Where possible, CDR templates are predicted by FREAD (Deane and Blundell, 2001; Choi and Deane, 2010) and multiple conformations of the hits are grafted as appropriate.

no prediction, *ab initio* modelling of CDRs using Sphinx (Marks et al., 2017) is used.

Side-chains are modelled by PEARS (Leem et al., 2018b), using a TCR-specific database. We used the STCRDab 2019 set as the starting point. Structures were clustered at 90% sequence identity. This resulted in a set of 123 structures, from which the database was constructed. See Leem et al. (2018b) for the full methodology.

When Sphinx is used, for performance reasons, we only remodel the side-chains of the loop grafted on by Sphinx, while keeping the rest of the side-chains unchanged.

3.2.9.3 Cross-validation set

To evaluate TCRBuilder, we carried out a leave-one-out cross-validation on 169 non-redundant TCR sequences in the STCRDab 2019 set, barring the algorithm from using full sequence-identical template structures.

3.2.9.4 Blind set

To test if TCRBuilder is comparable to current TCR modelling tools, we constructed a blind test set of the 10 (9 non-redundant) $\alpha\beta$TCRs deposited in the PDB between 24$^{\text{th}}$ July 2019 and 22$^{\text{nd}}$ October 2019. We tested LYRA and Repertoire Builder using their normal modes (Klausen et al., 2015; Schritt et al., 2019). TCRModel provides two modes for users. The results with no refinement are called "(norefine)", and the ones with refinement are referred to as "(refine)" (Gowthaman and Pierce, 2018). Table 3.2 lists the PDB codes tested in this "TCRBuilder blind set".

Table 3.2: List of PDB structures used in TCRBuilder blind set, and any structures that share identical sequences where applicable. PDB codes are listed as PDB ID followed by the TCR chain IDs.

PDB Codes (and structures with identical sequences)
6MSS_BA (6MRA_BA)
6MNM_BA
6MKR_BA
6MKD_BA
6JXR_nm
6MNO_BA
6MNN_BA
6MNG_BA
6Q3S_ED

3.2.9.5 Performance evaluation

TCRBuilder may produce more than one model if the CDR sequences are predicted to adopt more than one conformation. We compared the performance of the "conformer" model described below, to the native structure. While only one conformation was observed in a native structure, the predicted models might suggest CDR conformations that were previously unseen due to the low number of available TCR structures.

We assessed if any of the models in the ensemble captured a conformation that closely approximates the native structure. Hence, in the performance evaluation, we reported the lowest RMSD attained by the ensemble, to each of the native structures in the sequence-redundant set. The selected model is referred to as the "conformer" model. The backbone RMSD was calculated across the region of interest. For CDRs, the backbone RMSD was calculated after aligning the two residues immediately before, and two after, the given CDR loop.

3.3 Results

The IMGT-defined CDR loops (Lefranc et al., 2015) were extracted from each chain of the 270 high-quality TCR structures in the STCRDab 2018 set (see Section 3.2.3). This is more than double the numbers used in previous studies of TCR canonical forms, Al-Lazikani et al. (2000) (seven structures) and Klausen et al. (2015) (116 structures). Redundant sequences were retained as the conformations may differ as shown in the forthcoming sections. Summary statistics of the sequence-redundant STCRDab 2018 set for each of the CDR types are listed in Table 3.3.

Table 3.3: Summary of TCR and antibody CDR structural clusters in the STCRDab 2018 set and SAbDab 2018 set. "#" refers to the number of entities. TCR-only clusters are formed at a threshold of 1.0Å; TCR and antibody CDRs are clustered using the standard threshold for antibody CDRs given in Nowak et al. (2016). The TCR and antibody CDR sequence lengths observed in the combined TCR and antibody CDR cluster are separated by "/".

	TCR-only					**TCR and antibody**				
	α1	α2	α3	β1	β2	α1/L1	α2/L2	α3/L3	β1/H1	β2/H2
# Sequences	356	329	338	357	360	2817	3168	3210	3404	3328
# Unique sequences	45	44	99	33	39	625	242	954	826	924
Sequence lengths observed	5-8	5-8	8-15	5-6	5-7	5-8/3-12	5-8/2-7	8-15/5-14	5-6/4-15	5-7/6-12
# Canonical classes	7	7	5	1	2	13	3	11	9	7
# Pseudo-classes	3	3	11	1	1	15	14	36	23	17
# Sequences in canonical classes	318	248	126	344	340	2317	2856	2471	2868	2759
# Sequences in pseudo-classes	18	56	74	5	6	299	270	341	205	176
# Non-clustered sequences	20	25	138	8	14	201	42	398	331	393

3.3.1 Updating the canonical classes of TCR CDRs

We clustered each CDR type (*e.g.* CDRα1, CDRα2, etc.) using the method developed by Nowak et al. (2016) and described in Section 3.2.4. In total, we identified seven α1, seven α2, five α3, one β1 and two β2 canonical classes (see Table 3.3). The representative structures and sequence patterns of our TCR CDR canonical classes are shown in Figure 3.2 and Appendix Figures B.1 – B.4. We found that some canonical forms have highly conserved positions in their encoding sequences, which might govern the conformations. However, we also observed many cases where structures of the same sequence fell into different structural clusters. For instance, the α1 loop with the sequence DSVNN belonged to α1-5-A when

unbound, to pseudo-class α1-5-B* when bound to an MHC with a long peptide, and was non-clustered when bound to a short peptide (Figure 3.3).

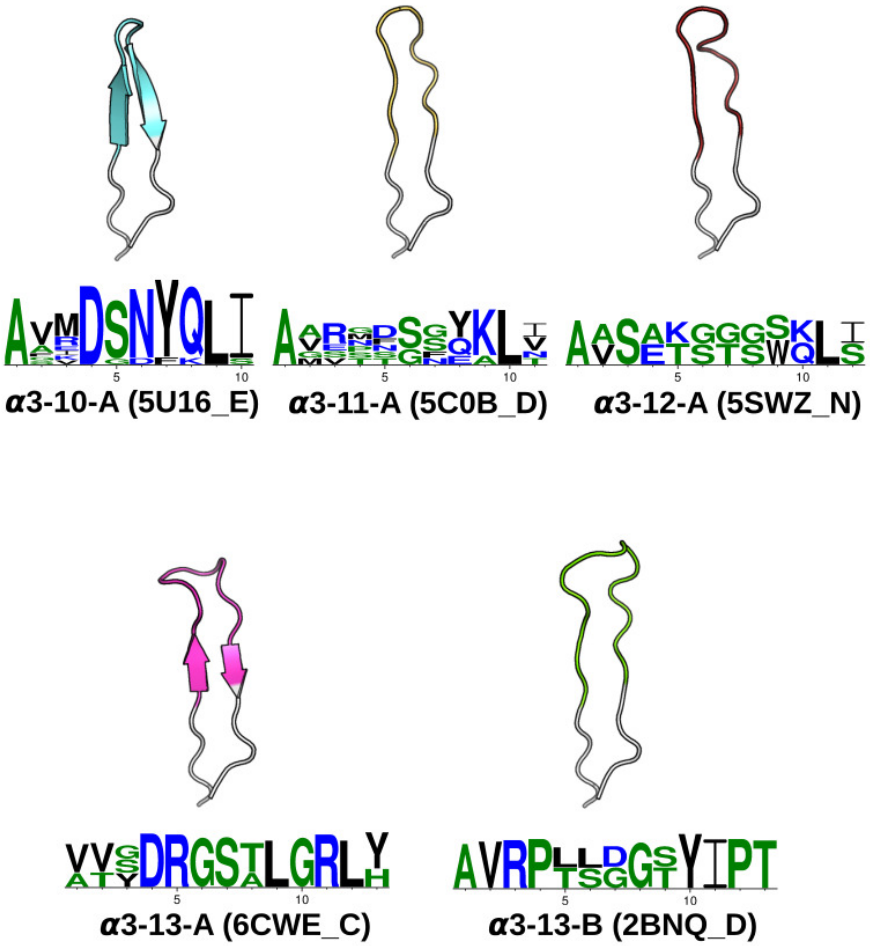


Figure 3.2: CDRα3 canonical classes. At a 1.0Å clustering threshold, our DBSCAN method identified five canonical classes. Each class has at least five structures and two unique sequences. For every CDRα3 class, the centroid structure is illustrated, with anchors in white, and the CDRα3 region (IMGT 105–117) coloured. The Protein Data Bank (PDB; Berman et al., 2000) four-letter code and the chain identifier of the centroid structure is shown in the bracket next to the cluster name. The sequence pattern below each centroid structure is generated by WebLogo (Crooks et al., 2004), using the unique sequences of the cluster: α3-10-A has 11, α3-11-A has 6, α3-12-A has 2, α3-13-A has 3 and α3-13-B has 2. Hydrophilic residues were in blue, neutral residues in green and hydrophobic residues in black (see Section 3.2.7).

Inter-species difference was not observed among canonical classes. Any antigenic species and TCR species may be found in any canonical classes (Appendix Figures B.5 and B.6). TCRs that bind to MHC1-related proteins appear to adopt

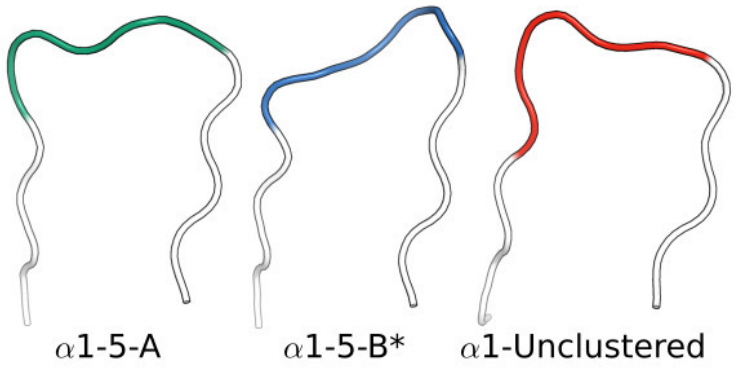


Figure 3.3: CDRα1 loops with the sequence DSVNN can form the α1-5-A canonical class (PDB: 5fka_A; unbound) or the α1-5-B* pseudo-class (PDB: 2ian_D; bound to MHC with a long peptide). Alternatively, they can also be non-clustered (PDB: 4mnq_D; bound to MHC with a short peptide).

only a particular set of CDR canonical forms (Appendix Figure B.7).

3.3.1.1 Comparison to previous canonical forms

We compared our canonical forms to those from Al-Lazikani et al. (2000) and Klausen et al. (2015) (see Tables 3.4 and 3.5). We matched canonical classes if their representative structure was found in our canonical class. Some canonical classes were represented by structures that were filtered out of our dataset for quality reasons. In these cases, a sequence-identical CDR with a comparable backbone conformation was used as a proxy to match canonical classes (see Section 3.2.5).

Table 3.4: Summary of CDRα canonical classes. The CDRα canonical classes from our study are compared to the canonical classes of Al-Lazikani et al. (2000) and Klausen et al. (2015) (see Section 3.2.5). Briefly, if there is at least one structure that is present in both our canonical class and those from the literature, we consider these classes to be analogous. Pseudo-classes are indicated with *.

Canonical class	**Al-Lazikani**	**Klausen**	**Number of unique sequences (Number of structures)**
α1-5-A		α1-1	7 (36)
α1-5-B*			1 (7)
α1-6-A	α1-1	α1-2	13 (106)
α1-6-B	α1-2	α1-3	6 (25)
α1-6-C		α1-4	5 (93)
α1-6-D			3 (18)
α1-6-E*			1 (5)
α1-6-F*		α1-5	1 (6)
α1-7-A		α1-7	6 (16)
α1-7-B		α1-6	3 (24)
(non-clustered)	α1-3		
α1-non-clustered			10 (20)
α2-5-A		α2-2	3 (16)
α2-5-B		α2-1	2 (22)
α2-6-A	α2-1	α2-3	4 (94)
α2-6-B*			1 (9)
α2-7-A	α2-2	α2-4	12 (56)
α2-7-B	α2-4	α2-5	8 (41)
α2-7-C*		α2-7	1 (28)
α2-8-A		α2-8	4 (9)
α2-8-B	α2-3	α2-6	3 (10)
α2-8-C*			1 (19)
α2-non-clustered			11 (25)
α3-9-A*			1 (5)
α3-10-A		α3-3	11 (53)
α3-10-B*			1 (12)
α3-10-C*			1 (8)
α3-10-D*			1 (5)
α3-10-E*			1 (6)
α3-11-A		α3-5	6 (27)
α3-11-B*			1 (7)
α3-11-C*		α3-6	1 (9)
α3-12-A		α3-7	2 (5)
α3-12-B*			1 (5)
α3-13-A		α3-8	3 (33)
α3-13-B		α3-10	2 (8)
α3-13-C*		α3-9	1 (6)
α3-13-D*			1 (6)
α3-13-E*			1 (5)

Continued on next page

Table 3.4 – continued from previous page

Canonical class	Al-Lazikani	Klausen	Number of unique sequences (Number of structures)
(filtered)		α3-1	
(non-clustered)		α3-2/α3-4/α3-11/α3-12	
α3-non-clustered			72 (138)

Table 3.5: Summary of CDRβ canonical classes. Clusters are labelled and mapped to previous canonical classes (see Table 3.4 and Section 3.2.5). If there is at least one structure that is present in both our canonical class and those from the literature, we consider these classes to be analogous. Pseudo-classes are indicated with *.

Canonical class	Al-Lazikani	Klausen	Number of unique sequences (Number of structures)
β1-5-A	β1-1/β1-2	β1-1/β1-2	30 (344)
β1-6-A*	β1-3		1 (5)
(non-clustered)		β1-3/β1-4	
β1-non-clustered			3 (8)
β2-6-A	β2-1	β2-2	18 (238)
β2-6-B		β2-3/β2-4/β2-5	16 (102)
β2-6-C*	β2-3	β2-6	1 (6)
(non-clustered)	β2-2	β2-1/β2-7	
β2-non-clustered			6 (14)

For CDRα1, our DBSCAN method broadly agreed with Al-Lazikani et al. (2000) and Klausen et al. (2015), apart from α1-3 in Al-Lazikani et al.'s classes, for which we found no corresponding cluster. Klausen et al.'s α1-5 cluster was matched to pseudo-class α1-6-F, meaning that for this cluster, we found more than five structures, but only a single unique sequence (Table 3.4).

All seven of our CDRα2 classes and our one CDRα2 pseudo-class were matched to previously observed CDRα2 canonical classes (see Table 3.4 for the full list of comparisons). All five of our CDRα3 canonical forms and two of the pseudo-classes mapped to ones from Klausen et al. (2015). In addition, we found nine further pseudo-classes that were not identified in their study. The cluster representative of the Klausen et al.'s α3-1 canonical class was filtered out of our dataset as it comes from a structure with a resolution greater than 2.9Å. Since its sequence

was also absent in the rest of our dataset, we were unable to map this class to our canonical forms. The cluster representatives of Klausen et al.'s α3-2, α3-4, α3-11 and α3-12 were non-clustered in our analysis.

Our single CDRβ1 class and our one CDRβ1 pseudo-class were matched to previous clusterings. Klausen et al.'s β1-3 and β1-4 forms were not in clusters in our work (Table 3.5). For CDRβ2, we were unable to find a match for Al-Lazikani et al.'s β2-2 class, nor Klausen et al.'s β2-1 and β2-7 classes. However, our two CDRβ2 classes and one CDRβ2 pseudo-class were matched, with our β2-6-B merging three of Klausen et al.'s clusters (Table 3.5).

Both previous clusterings were based on backbone dihedral angles of the CDR loops, whereas in our work, we clustered using backbone distances. Despite these different approaches, there was a large degree of overlap between our canonical forms and those found previously. We have also identified a small number of new canonical classes from our larger dataset. As was shown for antibody CDR canonical forms (Chapter 2; Wong et al., 2019a), the growth of structural data continuously modifies our understanding of CDR loop structures. It is therefore necessary to continuously and preferably automatically update the definition of canonical forms as more structural information becomes available.

3.3.2 Prediction of CDRs from sequence

As our TCR canonical classes showed conserved sequence patterns (Figure 3.2 and Appendix Figures B.1 – B.4), we built a sequence-based, length-independent PSSM, that can be used to predict TCR CDR canonical classes (Chapter 2; Wong et al., 2019a). The performance of the predictor was evaluated by a leave-one-out cross-validation protocol on the unique sequences in STCRDab 2018 set (Table 3.6). Accuracy ranged from 73.2% to 100%, which is comparable to previous results (Klausen et al., 2015).

To assess the potential of our method on large sets of sequencing data, we used our PSSMs to predict the CDR canonical classes of an Ig-seq dataset of mouse

Table 3.6: Leave-one-out cross-validation accuracy.

CDR	Unique sequences	PSSM Accuracy
CDRα1	44	81.8%
CDRα2	43	73.2%
CDRα3	91	76.1%
CDRβ1	32	100%
CDRβ2	38	92.1%

TCRα sequences (Maceiras et al., 2017). The entire dataset contained 1,563,876 sequences (1,498,254 CDRα1, 1,563,876 CDRα2, 1,267,235 CDRα3); on a single 3.4GHz core, the prediction took three minutes. Our method achieved high coverage for CDRα1 and CDRα2, but made predictions for only 37% of the non-redundant CDRα3 sequences (Table 3.7). The poor coverage of CDRα3 could be due to the paucity of the currently available data in capturing the conformations of this more sequence-diverse CDR.

Table 3.7: Prediction of CDRα sequences from Tfh and Tfr cells.

CDR	Prediction of redundant sequences	Prediction of unique sequences
CDRα1	283940/292310	1084/1278
CDRα2	313317/313546	1359/1435
CDRα3	116246/232260	1520/4139

Ten TCR structures containing 44 individual chains that were unseen at the time of methodology development and before 5th June 2019, were used as a blind test set for the predictor in the SCALOP-TCR blind set (see Section 3.2.6.5, Table 3.1). All canonical CDRs had 100% prediction coverage. Apart from CDRα3, all CDR types were predicted with 100% accuracy; in other words, at least one member of the predicted canonical class had backbone root-mean square deviation (RMSD) $\leq$1.0Å to the native structure. CDRα3 had one false prediction where the loop GTERSGGYQKVT was not assigned to any clusters even though the backbone RMSD falls within 1.0Å to a member of the α3-9-A canonical class.

3.3.3 Comparison between TCR and antibody CDRs

Inspired by the shared architecture and genetic generation mechanism of antibodies and TCRs, we examined whether their CDR length distributions, structures/canonical forms, and sequence patterns that give rise to their canonical forms overlap. For each of the TCR CDRs, we compared all its structures with the corresponding antibody CDR's structures (*e.g.* CDRα1 with CDRL1). We considered the CDR sequence length distributions and their structural clustering with DBSCAN, using the clustering thresholds that have been previously used for antibody CDR clustering (Nowak et al., 2016). Changing these thresholds does not qualitatively affect the results described below. A full list of comparisons is described in Table 3.3. For all pairs of TCR/antibody CDRs, we found that most clusters contained only TCR CDRs or antibody CDRs (*e.g.* CDRα1 only). In other words, the CDRs from TCRs and from antibodies tended to occupy distinct areas of structural space. Where we observed an overlap in the structural space, we inspected whether the sequence motifs for the canonical classes differ between the two types of receptors.

3.3.3.1 CDRα1/CDRL1

The CDRα1 loops in our set were five- to seven-residues long, while CDRL1 loops spanned from three- to twelve-residues (Appendix Figure B.11). The most common length was six for both types of loops: over 70% of the CDRα1 loops and 45% of the CDRL1 loops. Thirteen structural clusters of CDRα1/CDRL1 were identified (Appendix Figure B.12); of these only α1,L1-5-A contained both CDRα1 loops (6 unique sequences) and CDRL1s (18 unique sequences). The sequence logos formed by the sequence-unique CDRα1 and CDRL1 loops in α1,L1-5-A had different sequence patterns (Figure 3.4). The general physicochemical properties were similar, but CDRL1 had a preference for Valine and Tyrosine on the third and fifth positions while CDRα1 showed ambiguity at these two positions. Principal component analysis (PCA) on one-hot-encoded unique sequences displayed a separation between CDR loops from the two types of receptors (see Figure 3.4), except for one TCR α1

sequence (DSVNN) that was considered to have a more similar sequence pattern to antibody L1 loops. This sequence is a TCR α1 sequence that adopts multiple conformations as described above.

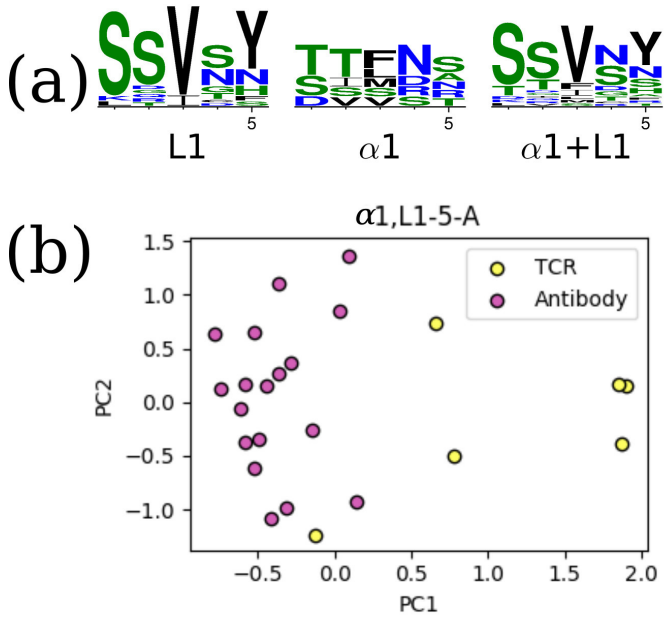


Figure 3.4: The unique TCR and antibody sequences in the α1,L1-5-A class. (a) Sequence logos of CDRα1 and CDRL1 loops in the α1,L1-5-A class, using only the unique sequences. The sequence patterns are distinct between the two classes. (b) Principal component analysis (PCA) plot of the first two components in one-hot-encoded sequences, stratified by TCR and antibody CDRs (see Section 3.2.7). TCR α1 and antibody L1 can be separated by the first principal component. One TCR sequence (DSVNN) is close to the antibody set.

3.3.3.2 CDRα2/CDRL2

CDRα2 loops tend to be longer than CDRL2, with nearly all of the CDRL2 loops being three-residues long and CDRα2 having a range of lengths from five to eight (Appendix Figure B.13). Given this, there is little chance of structural similarity between the two types of loops. Structural clustering confirmed that all classes only contained one CDR type, with two CDRα2 clusters, and one CDRL2 cluster (Appendix Figure B.14).

3.3.3.3 CDRα3/CDRL3

The majority of the CDRL3 loops in our set were nine-residues long (2073 of the 2872; 72.2%), whilst CDRα3 loops had nine to fifteen residues (Figure 3.5). Among the eleven clusters identified, we found a cluster, α3,L3-10-A, which included both CDR types, with 17 and 12 unique sequences of CDRL3 and CDRα3 respectively (Figure 3.6). A PCA of the sequences in this cluster (α3,L3-10-A) separated the TCRα3 and antibody L3 members with one exception (Figure 3.7). The α3 sequence GTYNQGGKLI clustered with the L3 group. This sequence was one of the eight (out of a total of 99) unique α3 sequences we found to have multiple conformations. Our dataset contained three structures of this sequence; the other two were non-clustered (3VXU_I and 3VXU_D).

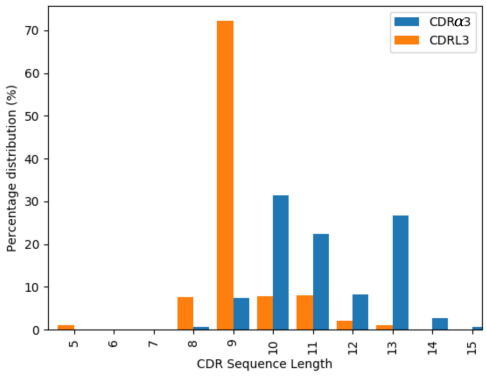


Figure 3.5: Length distributions of CDRα3 (blue) and CDRL3 (orange) loops.

3.3.3.4 CDRβ1/CDRH1

In our dataset, CDRβ1 loops tended to be shorter than their CDRH1 counterpart (Appendix Figure B.15). Over 80% of the CDRβ1 were five-residues long, compared to eight-residues long loops dominating for CDRH1 (>80%). Nine classes were formed by the structural clustering: seven CDRH1 clusters and two CDRβ1 clusters (Appendix Figure B.16).

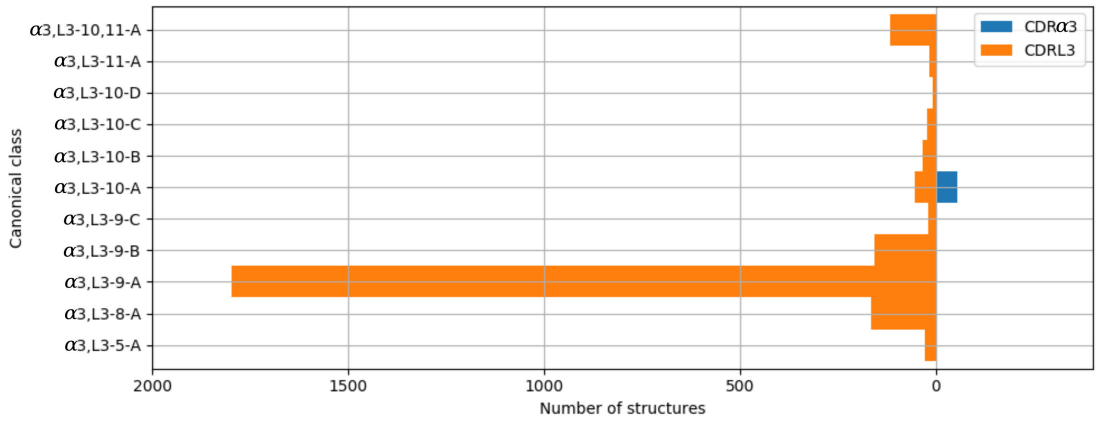


Figure 3.6: Clusters of CDRα3 (blue) and CDRL3 (orange) loops. All CDRα3 and CDRL3 structures were clustered using the DBSCAN method. Orange bars indicate the number of CDRL3 structures, and blue bars the number of CDRα3 structures. All classes, apart from α3,L3-10-A, have structures from only one CDR type, *i.e.* CDRα3 or CDRL3.

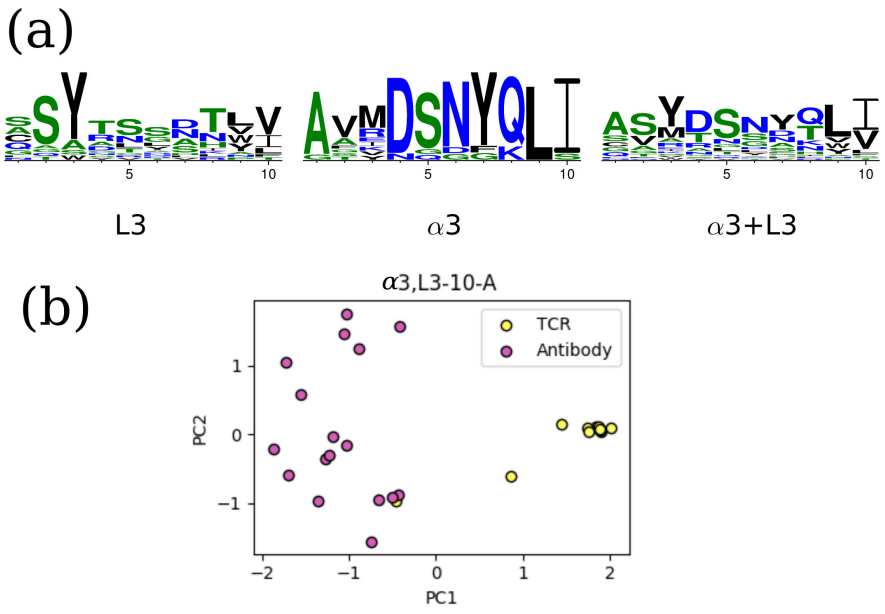


Figure 3.7: The unique TCR and antibody sequences in the α3,L3-10-A class. (a) Sequence logos of CDRL3 and CDRα3 loops in the α3,L3-10-A class, using only the unique sequences. The sequence patterns appear to be distinct between the two classes. (b) Principal component analysis (PCA) plot of the first two components in one-hot-encoded unique sequences in α3,L3-10-A, stratified by TCR and antibody CDRs (see Section 3.2.7). TCR α3 and antibody L3 are separated by the first principal component, only one TCR sequence (GTYNQGGKLI) is close to the antibody set.

3.3.3.5 CDRβ2/CDRH2

More than 90% of the CDRβ2 loops in our set were six-residues long, while only 0.2% of the CDRH2 had six residues and the rest were in the range of seven to ten residues (Appendix Figure B.17). None of the seven clusters that were formed contained members from both types of CDRs (Appendix Figure B.18).

3.3.3.6 β-chain CDRs in TCRs and heavy-chain CDRs in antibodies and nanobodies

In our clustering, CDRs from nanobodies were also included. Unlike the CDRs of TCRs, these typically did fall into the antibody H1 and H2 clusters (Appendix Figures B.16 and B.18). In the case of CDRβ1/CDRH1, 72.2% (156 out of 216) of the nanobody CDRH1 loops were clustered with antibody CDRH1s in the β1,H1-8-A class. The remaining nanobody CDRH1 loops formed the majority of the other two length-eight clusters (all 46 in β1,H1-8-B and 14 out of 21 in β1,H1-8-C respectively). Nanobody CDRH2 loops clustered with the length-seven and eight antibody CDRH2 loops. These results suggest that nanobody CDR loops are structurally more similar to antibodies than to TCRs.

3.3.3.7 CDRβ3/CDRH3

There are no canonical classes for CDRH3 or CDRβ3 but in the case of CDRH3, previous studies have found structural conservation in the start and end of the loop, known as the "torso" or "base" region (Weitzner et al., 2015; Finn et al., 2016). We therefore inspected the base structure of the CDRβ3/CDRH3 loops. Following the study by Weitzner et al. (2015), we carried out the loop anchor transform (LAT) analysis that was used to capture the characteristic base structure in CDRH3. The LAT analysis showed that CDRβ3 loops have a similar width of distribution in all six degrees of freedom as seen for CDRH3, but with a slight shift in the peak (Appendix Figure B.19). This presented the possibility that CDRβ3 and CDRH3 could share similar base structures. In the same study, the pseudo bond angle (τ) and pseudo dihedral angle (α) of the penultimate residue of the CDRH3 structure (IMGT position 116) were used to differentiate between extended

and kinked torsos. We found that very few CDRβ3 loops had their τ_{116} and α_{116} in the space of kinked torsos (Figure 3.8). Instead, 317 out of the 325 (97.5%) CDRβ3 loops had an extended base. This behaviour is the opposite of CDRH3 loops. The eight outlying CDRβ3 structures with positive α_{116} have either an aromatic side chain at IMGT position 116 that restricts the shape of the base, or an abrupt bend to accommodate unusual binding peptides. Consistent observations were made when we analysed the ϕ/ψ plots for the torso positions as outlined in Finn *et al.* (2016; see Appendix Figure B.20).

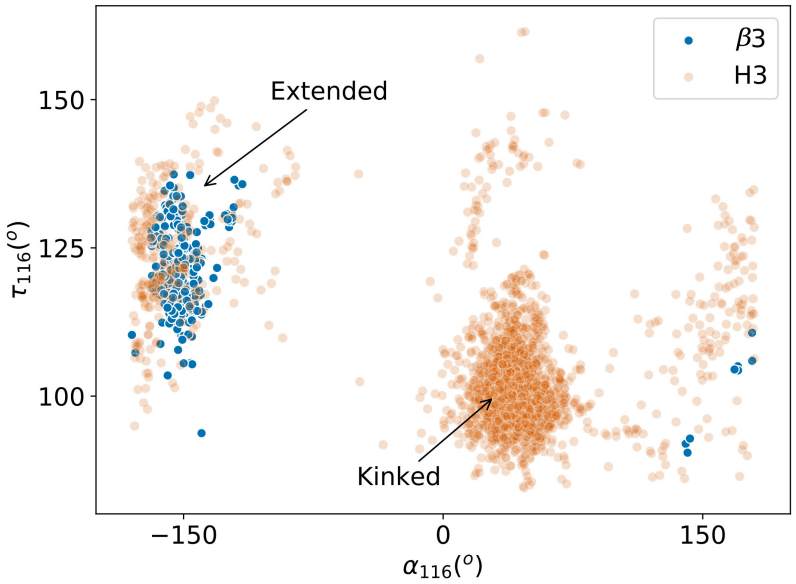


Figure 3.8: Pseudo bond angle (τ) and pseudo dihedral angle (α) analyses on IMGT (Lefranc et al., 2015) position 116. Scatters represent individual observations of β3 (blue) and H3 (orange) structures. Regions occupied by kinked and extended CDRH3 loops according to Weitzner et al. (2015) are indicated by arrows.

3.3.4 CDR structural variability in TCR and antibodies

As noted above, some TCR CDRs with identical sequences fell into different structural clusters. Considering the joint clustering of TCR and antibody CDRs, we found that of the 212 unique TCR CDR sequences with multiple example structures, 41 (19.3%) fell into more than one structural cluster (Appendix Table B.1). This

happened far less for antibody CDRs where only 7.95% (166 of 2089) of the unique antibody CDR sequences with multiple example structures were found in more than one structural cluster. Whether the structure was crystallised in complex with the cognate molecule did not appear to cause the conformational difference as bound and unbound structures were observed in the different clusters (Appendix Tables B.2 and B.3).

Following the analysis in Marks and Deane (2017), we compared the maximum backbone RMSD of structures from an identical sequence (Figure 3.9). If we consider structures with a difference of less than 1Å in backbone RMSD as similar (184 of the 284 TCRs and 2195 of the 2717 antibodies), this analysis illustrates that the majority of the sequence-identical structures are structurally similar in both TCR and antibody CDRs, but that TCR loops show a more skewed distribution, tending towards higher structural variability.

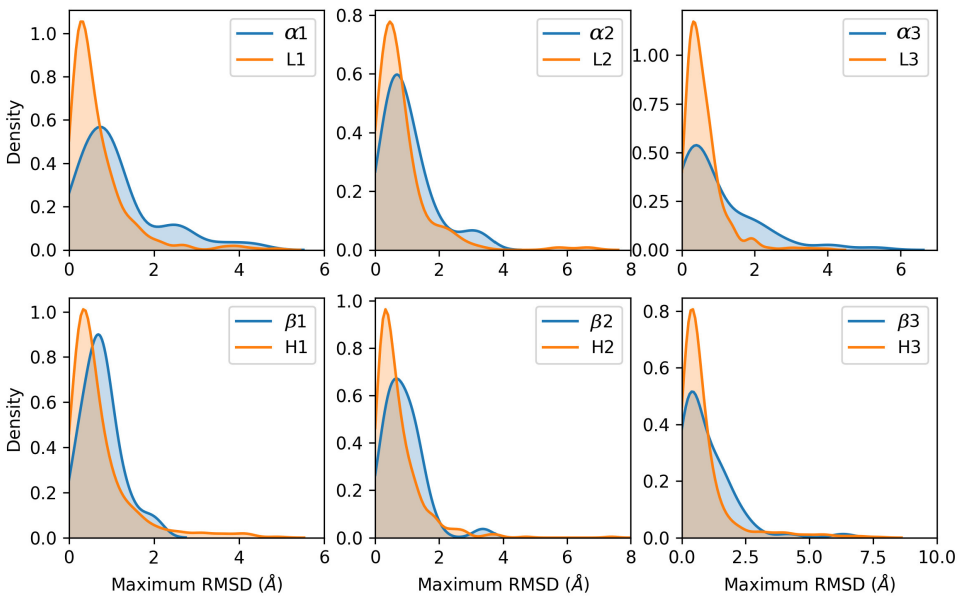


Figure 3.9: The density of the maximum RMSD between loop structures with identical CDR sequence. TCR CDRs (blue) show a shift to the right compared to antibody CDRs (orange), suggesting that structural flexibility of TCR CDRs is greater than that of antibodies.

3.3.5 Multi-state TCR homology modelling

The higher structural variability in TCR CDRs than in antibody CDRs has not been captured by current TCR modelling tools. To more accurately reflect the multi-state CDR conformations in TCR, we built a homology modelling tool, TCRBuilder, that returns an ensemble of TCR models if the CDRs were predicted to be in multiple conformations.

3.3.5.1 Performance analysis

To evaluate TCRBuilder, we carried out a leave-one-out cross-validation on the 169 non-redundant TCR sequences in the STCRDab 2019 set (see Section 3.2.9.3). The average RMSD for the complete model was 1.49 Å. In the models, CDRs 1 and 2 on both the α and β chains are well predicted, but CDR3 less so (Figure 3.10). Of the 169 models generated, 106 (62.7%) required template-free modelling (Marks et al., 2017) for at least one of their CDRs. This is a far larger percentage than the <4% of antibody cases that required template-free modelling in a large test (Leem et al., 2016). These results highlight the lack of structural data available for TCRs.

To test if TCRBuilder is comparable to current TCR modelling tools, we used the TCRBuilder blind set of the 9 non-redundant $\alpha\beta$TCRs (see Section 3.2.9.4). We used LYRA and Repertoire Builder in their normal modes, and both the "refine" and "norefine" modes of TCRModel were tested. Figure 3.11 shows that TCRBuilder has comparable performance to the other four TCR modelling web servers.

3.3.5.2 Predicting multiple conformations

As TCR CDR sequences are structurally variable, TCRBuilder is designed to capture the distinct conformations of CDRs that are known to exist. Eighty-nine of the 169 evaluated models were predicted to have more than one conformation of their binding sites. Figure 3.12 shows an example of a CDRα3 sequence that can have

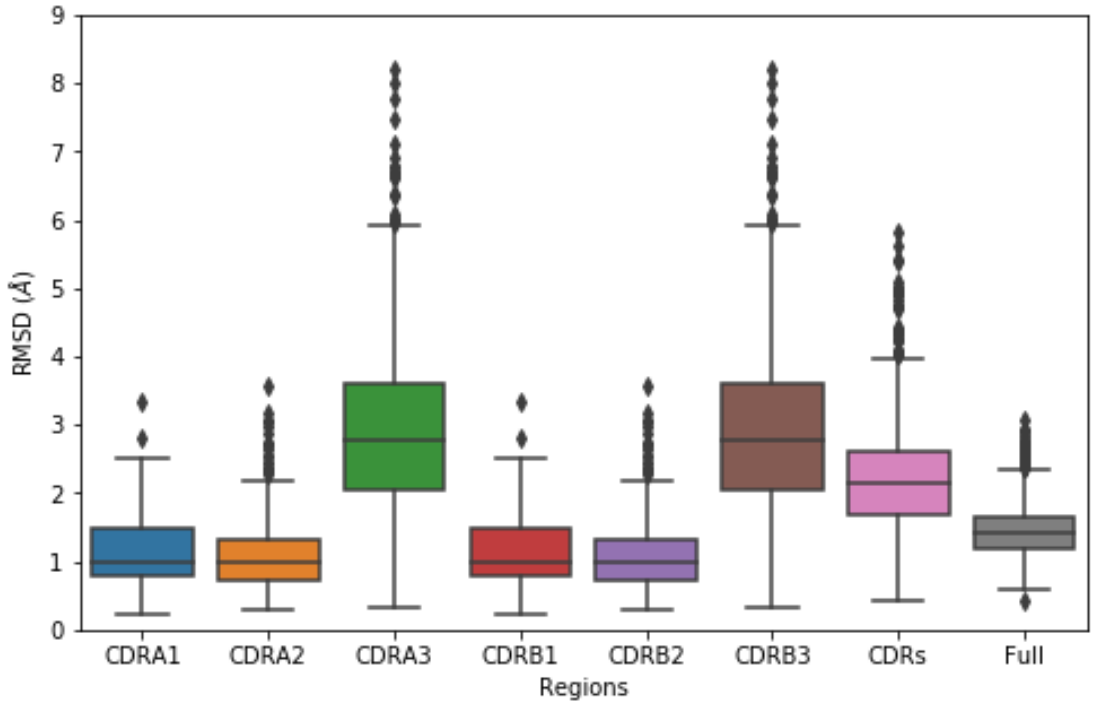


Figure 3.10: Cross-validation performance. The root-mean-square-distance (RMSD) of all backbone atoms across the respective regions, between the native structure and the model with the lowest RMSD. The RMSD for "CDRs" is calculated by dividing the backbone RMSD in all CDRs divided by the total lengths of all CDR sequences; whereas the RMSD for "Full" is calculated over the framework and CDR regions.

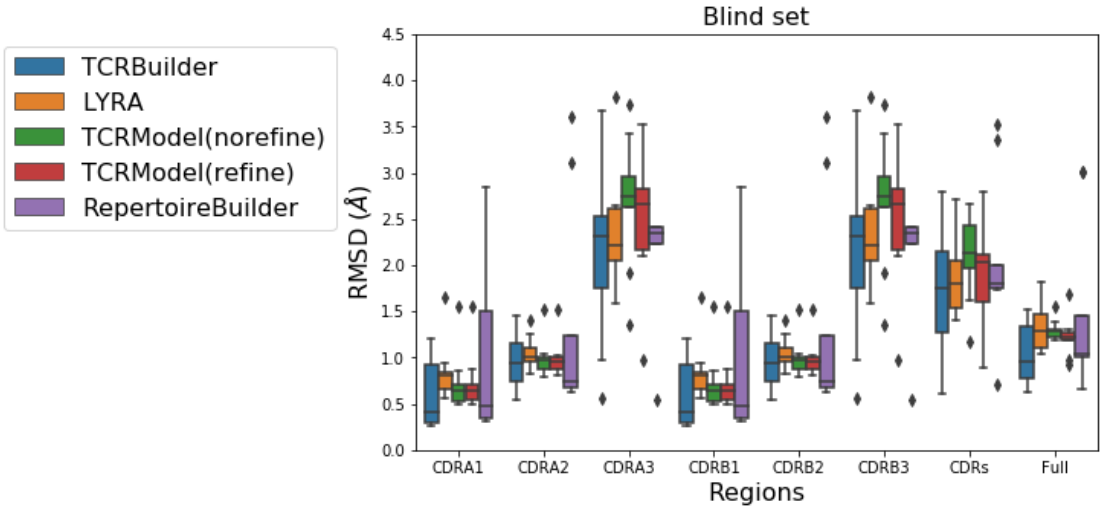


Figure 3.11: Benchmark performance on the TCRBuilder blind set with other existing TCR modelling web servers. TCRBuilder has comparable performance to the existing tools.

multiple distinct conformations. As the number of TCR structures increases, it is likely that we will find many more CDR sequences with variable conformations.

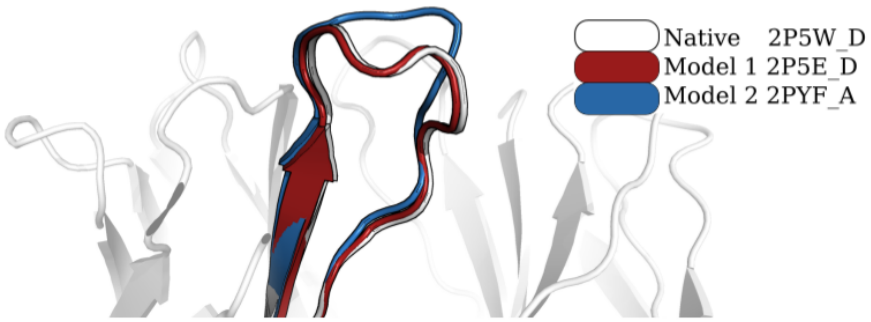


Figure 3.12: An example case where a CDRα3 sequence (AVRPLLDGTYIPT) with multiple distinct conformations was modelled. One of the conformations (2P5E_D in red) was nearly identical to the native 2P5W_D structure, but an alternative conformation of this loop (2PYF_A in blue) is also predicted as a possible shape.

3.4 Discussion

We have identified and refined the canonical class definitions of TCR CDRs using a more up-to-date structural dataset, and used these to generate an auto-updating database and prediction server. One of the major advantages of a canonical class model is the rapid mapping between the sequence and structural space. To demonstrate this for TCRs, we applied an adapted PSSM-based methodology (Chapter 2; Wong et al., 2019a) to assign the canonical classes for an Ig-seq dataset of $\sim$1.5 million mouse TCRα sequences in a few minutes.

The commonalities between TCR and antibody in their folds and the genetic mechanisms that generate them prompted us to explore the similarities and differences between TCR and antibody CDR structures. We performed a length-independent structural clustering of TCR and antibody CDRs and found that they almost always separate into different clusters. This separation was partly driven by the differences in the length distributions between TCR and antibody CDRs. In the cases where we found structural clusters which contained both types of loops, we found that the underlying sequence motifs were distinct.

We also compared the CDRβ3 and CDRH3 loops in terms of their torso structures – the starting and the ending residues of the CDRβ3/CDRH3 loops, and found once again TCRs and antibodies gravitated towards different structures. CDRH3 loops in antibodies tend to have a kinked torso while CDRβ3 are only found with the extended torsos that is less common in CDRH3.

Driven by observations that many sequence-identical TCR CDR structures were found in different clusters, we also assessed the structural variability of TCR and antibody CDRs. Our results agree with previous findings about the flexibility of TCR CDRs (Holland et al., 2018; Crooks et al., 2018), and further suggest that TCR loops are more flexible than antibody CDRs. Structural rigidification has been proposed as an affinity maturation mechanism in antibodies (Mishra and Mariuzza, 2018), while flexibility in the binding site has been suggested as one of the features enabling the promiscuous binding of TCR-MHC (Rossjohn et al., 2015). The fact that nearly 20% of the TCR CDRs that could show different conformations did so suggests that the canonical class model will struggle to accurately predict TCR CDR conformations. TCR modelling tools may provide more details through homology or *de novo* loop predictions (Klausen et al., 2015; Gowthaman and Pierce, 2018; Li et al., 2019). Overall, our analyses suggest that there are structural differences between TCR and antibody CDRs, and the differences we observe potentially help explain how these two receptors bind to their separate target antigens.

To this end, we built TCRBuilder, a TCR modelling server that reflects the high structural variability in TCR binding sites. The multi-state models it generates may better represent the possible binding configurations a TCR can adopt. This could be used to seed multiple starting points for molecular dynamic simulations, or suggest multiple possible binding poses in TCR-MHC complex modelling.

Many groups have attempted to augment TCR and antibody designs by swapping binding sites between these two receptors (*e.g.* Xu et al., 2018, 2019). TCR-mimic antibodies were developed in the hope of transferring the ability of TCRs to target

intracellular proteins, to antibodies for the use in cancer therapy. Antibody-like TCR design transfers the highly specific antibody binding sites to TCRs. It was shown to enhance the specificity and affinity in TCR antigen recognition *in vitro* and *in vivo* (Xu et al., 2018). These cases present the possibility of altering the behaviour of the receptors such as targeting peptides in the context of an MHC molecule by grafting TCR binding sites, or enhancing specificity when the antibody components are added. Our results should aid the development and design of these types of therapeutic immune proteins.

3.5 Chapter Summary

In this chapter, we have described the comparative analysis of CDR structures in TCRs and antibodies, and a TCR structural modelling tool, TCRBuilder, that is able to capture the ensemble of CDR conformations in TCRs.

We first updated the canonical class definitions of TCR CDRs. Adapting the PSSM-based predictor presented in Chapter 2, we built a sequence-based TCR CDR canonical form predictor (SCALOP-TCR) that could be used to enrich sequencing datasets with structural information.

The genetic mechanisms and architecture between antibody and TCR are highly similar, but their biological functions differ. To understand the features that differentiate their biological functions, we studied the structural properties in their binding site. By comparing the antibody and TCR CDRs, we found that they tend to have different lengths. Where they shared similar lengths, our structural clustering of CDRs suggested that they fell into distinct structural spaces. In the two structural clusters where both antibody and TCR CDRs were found, we observed that antibody and TCR CDRs used different sequence patterns. We ran an anchor analysis of the highly variable CDRβ3/CDRH3 loops, and showed that

CDRβ3 loops tended to have an extended torso that was less common in CDRH3.

Furthermore, the structural variability in TCR CDRs was much higher than that in antibody CDRs – almost 20% of TCR CDRs that could adopt multiple conformations did so, compared to <8% of antibody CDRs. This observation suggests that an ensemble approach for TCR structural modelling would be highly beneficial.

Our informatics analysis complements previous experimental studies on TCR CDR structural variability (*e.g.* Rossjohn et al., 2015) and highlights some of the key considerations when using antibody to design TCR, or *vice versa*. The multi-state structural modelling provides a method to study the promiscuous binding in TCRs to peptide-MHC complex. The models can collectively represent alternative binding modes for the TCR-peptide-MHC engagement, paving the way to understand polyspecificity in TCRs.

In the next chapter, we will return to antibody binding site analysis. We will introduce Ab-Ligity, an antibody binding site comparison method that can identify sequence-dissimilar antibodies that bind to the same epitope. Ab-Ligity is adapted from its small molecule counterpart, Ligity (Ebejer et al., 2019), and considers the spatial separation of residues by their physicochemical properties.

4

Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope

Contents

This chapter is adapted from a preprint (Wong et al., 2020b).

4.1 Introduction

The previous chapters covered an analysis of the backbone conformations of CDRs. In this chapter, the overall binding site structure is considered to evaluate the similarity between two binding sites.

The highest resolution method for studying antibody-antigen binding configurations is co-crystal complex structures. These give atomic level information but are expensive and difficult to obtain (Dunbar et al., 2014b). Experimental mapping is often used as a surrogate as it is able to identify the binding regions

(see Section 1.4.1 for more details). Competition assays exploit the cross-blocking effect of antibodies that displace one another if they bind to similar or neighbouring epitopes (Kwak and Yoon, 1996; Abdiche et al., 2017). This method gives a coarse representation of which binders may share similar target sites, as minimal epitope overlap is sufficient for a pair of antibodies to compete with each other (Abdiche et al., 2017). A more refined approach is hydrogen deuterium exchange (HDX). HDX assesses the solvent accessibility of the bound and unbound forms of the partner proteins, and highlights regions with the maximum changes upon binding (*e.g.* Zhang et al., 2018; Puchades et al., 2019). The resolution is typically up to the range of peptides in the immediate proximity of the binding site. To achieve residue-level resolution, point mutations of the interacting proteins can be used to indicate the key binding residues. Mutagenesis studies measure the binding kinetics upon mutation of specific residues, but structural integrity may be compromised by the mutations, leading to spurious results (Abbott et al., 2014). All three of these experimental techniques provide an approximation of the binding regions, but are usually unable to provide a fine mapping of exact epitopes and paratopes.

Computational techniques have also been developed to identify antibodies that bind in similar ways. These have generally been exploited on large immunoglobulin sequencing datasets (*e.g.* Galson et al., 2015). Many of these techniques require a large number of known binders to a given epitope, to be able to identify further binders (Mason et al., 2019). The informatics approaches employed to analyse datasets, when none or only a few binders to a given epitope are known, are mainly dependent on sequence similarity. This is based on the concept of "clonotype" analysis (Galson et al., 2015), which considers the genotype and the sequence identity of CDRH3 (see Section 1.3; Galson et al., 2015; Trück et al., 2015; Soto et al., 2019). Clonotyping exploits the evolutionary origins of antibodies, using both the concept that antibodies from the same clone and lineage tend to bind similarly (Hsiao et al., 2020), and that the CDRH3 region, the most sequence-variable region, is often responsible for much of the binding (North et al., 2011).

Whilst clonotype analysis can identify antibodies with similar binding modes, there are many cases where binding remains the same even when the CDRH3 sequences and/or the genotypes are different (*e.g.* in anti-lysozyme antibodies; Pons et al., 2002).

To capture potential chemical interactions and the antibody-antigen binding configurations, macromolecular docking has been used (Lensink et al., 2019). Some studies have also presented the possibility of enriching computational modelling with experimental information (*e.g.* Cannon et al., 2019). However, both these methods are slow and not scalable to the large datasets of antibody sequences available in the early discovery stage (Raybould et al., 2019b).

In small molecule discovery, comparing the spatial arrangements of pharmacophores (the atom features involved in interactions) has been proposed as an alternative to docking when searching for similar binders. Pharmacophoric points are encoded using geometric hashing algorithms (*e.g.* Shulman-Peleg et al., 2005; Wood et al., 2012; Ebejer et al., 2019; Section 1.4.2.2). This approach is much faster than docking whilst giving comparable results (Ebejer et al., 2019). Some of the descriptors for ligand-binding pockets have been adapted for protein-protein interactions. An example is I2ISiteEngine, that uses triangulation to describe protein surfaces, albeit at a low throughput (Shulman-Peleg et al., 2004). More recently, geometric deep learning-based MaSIF (Gainza et al., 2020) and the simulated annealing algorithm InterComp (Mirabello and Wallner, 2018) have been proposed for comparing general protein-protein interfaces and tested on solved crystal structures, not on models and more specifically, not on models of antibodies. In the context of finding antibodies in large sequencing datasets that target the same epitope, a tool needs to be fast, applicable on antibody-antigen interfaces, and able to cope with the predicted binding site structures of antibodies.

In this chapter, we describe Ab-Ligity, an antibody version of the small molecule method Ligity (Ebejer et al., 2019). Ab-Ligity uses residue points tokenised by their physicochemical properties to allow rapid operation. To simulate a real-life scenario, we applied Ab-Ligity to antibody models and predicted paratopes, and showed that it can accurately predict antibodies that share similar target epitopes. As the majority of the similar binders shared similar sequence composition and lengths, and can be identified by sequence-based metrics, we removed these easy matches. In these more challenging cases of sequence-dissimilar antibodies, Ab-Ligity still accurately predicted antibodies that have similar binding modes. Ab-Ligity also performed better than InterComp, in a fraction of the time. Finally we described two case studies where Ab-Ligity predicts sequence-dissimilar and length-mismatched antibodies that bind to highly similar epitopes.

4.2 Method

4.2.1 Antibody-antigen co-crystal datasets

We selected all paired antibody-antigen complexes from SAbDab (Dunbar et al., 2014b) as of 27$^{\text{th}}$ January 2020, that were solved by X-ray crystallography, had no missing residues in all six CDRs (using the North definition; North et al., 2011), and were co-crystallised with a protein antigen of more than 50 residues. To avoid redundancy, we retain only one copy of each antibody. In the case of multiple copies, the complex with the best resolution is selected. Nine hundred and twenty antibody-antigen complexes were identified.

We defined epitopes as residues on the antigen with any atoms within 4.5Å of its cognate antibody. We extracted these residues using Biopython (Cock et al., 2009).

4.2.2 Antibody modelling and paratope prediction

We numbered antibody sequences using the IMGT scheme (Lefranc et al., 2015) and defined the CDR regions by the North scheme (North et al., 2011), with ANARCI (Dunbar and Deane, 2016). We then modelled the full set using ABodyBuilder (Leem et al., 2016), barring it from using sequence-identical structures as templates. The ABodyBuilder template library was built using all structures available on the 27$^{\text{th}}$ January 2020 in SAbDab (Dunbar et al., 2014b).

We used Parapred to predict the paratope residues (Liberis et al., 2018). Parapred gives a score to each residue in the CDRs, and two residues before and after, to indicate how likely it is to participate in binding. As suggested in the original paper, we selected residues with a score of $\geq$0.67 as the predicted paratope residues on the models (see Appendix C.3 for Parapred prediction performance). The coordinates of these predicted paratope residues were obtained from the corresponding model. Models and predicted paratopes were used throughout the chapter for calculating the structural similarity between antibodies.

4.2.3 Ab-Ligity calculations

The workflow of Ab-Ligity is illustrated in Figure 4.1. The binding residues are tokenised according to Figure 1.4 and Table 4.1. This tokenisation scheme was chosen as it is simple and intuitive, and gave similar results to several other more complex choices. For each binding site residue, a point is placed representing the residue group on the C_α atom of the residue. These points collectively represent a paratope on an antibody or an epitope on an antigen.

Ab-Ligity uses the hashing function outlined in (Ebejer et al., 2019). For a paratope or an epitope, it considers all combinations of triplets formed from a set of tokenized residues in a binding site. In a given triplet, each edge is represented by its vertices' tokens and its length. Each combination of tokens has a unique hash code.

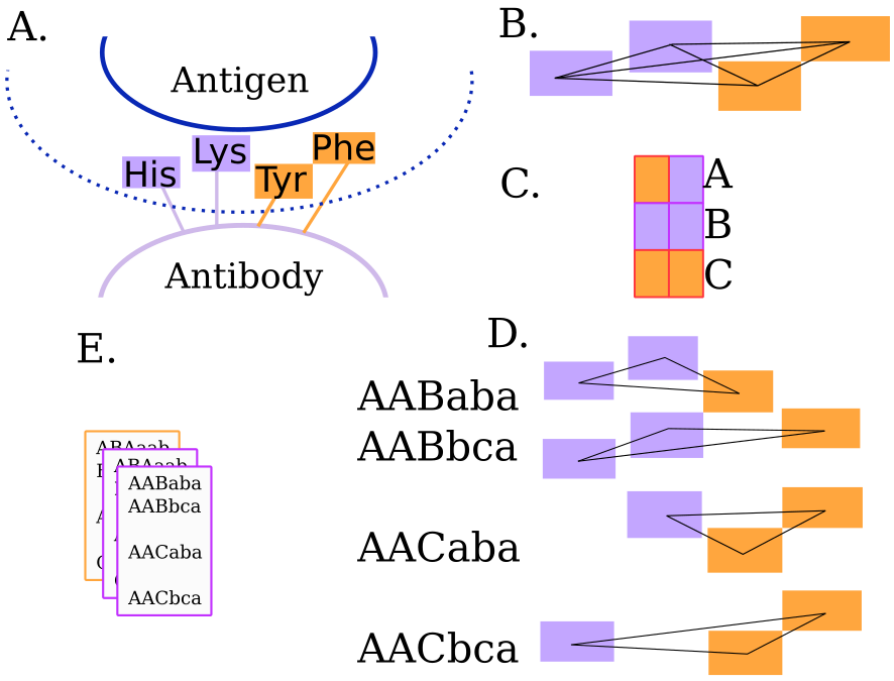


Figure 4.1: The Ab-Ligity workflow. **A.** Binding site residues within 4.5Å of the binding partner are tokenised as stated in Figure 1.4 and Table 4.1. **B.** All distances between tokenised points are calculated and hashed into 1.0Å-wide distance bins for both the paratopes and epitopes. **C.** Each pair of tokens is given a unique hash code. **D.** A six-character hash code is generated for each triplet. **E.** The hashes of a binding site are stored in a frequency table.

Table 4.1: Residue groupings for tokenisation.

Group	Residues
Aliphatic	Glycine (G), Alanine (A), Valine (V), Leucine (L), Isoleucine (I), Proline (P)
Hydroxyl	Serine (S), Threonine (T)
Sulphur	Cysteine (C), Methionine (M)
Aromatic	Phenylalanine (F), Tyrosine (Y), Tryptophan (W)
Acidic	Aspartic acid (D), Glutamic acid (E)
Amine	Asparagine (N), Glutamine (Q)
Basic	Histidine (H), Lysine (K), Arginine (R)

For instance, Aliphatic-Aliphatic would have a hash code of "a", Aliphatic-Aromatic and Aromatic-Aliphatic both have a hash code of "b" and so on. Edge lengths are discretized into bins of 1.0Å in both paratopes and epitopes to reduce computational complexity: edges between 1.0Å and 2.0Å are put in the bin A, between 2.0Å and 3.0Å in bin B and so on. The final hash code for a given triplet is determined by the three vertex hash codes sorted alphabetically, followed by the hash codes of the three corresponding length bins. The resulting hash table for the target binding site stores the frequency of these hash codes from all triplets.

Binding site similarity is calculated by the Tversky index of the pairs of hash tables:

$$S(X,Y) = \frac{|X \cap Y|}{|X \cup Y| + \alpha|X - Y| + \beta|Y - X|} \tag{4.1}$$

where X and Y are the two hash tables, $|X \cap Y|$ and $|X \cup Y|$ are the intersection and union, and $|X - Y|$ and $|Y - X|$ are the differences between the two tables respectively. For this study we used $\alpha = \beta = 0.5$.

4.2.4 Performance evaluation settings

We calculated the Ab-Ligity paratope similarity based on antibody models and their predicted paratopes. The majority of the antibodies that engage the same epitope have highly similar CDRH3 sequences and would be predicted to bind in the same manner by sequence comparison (such as clonotype). To test if Ab-Ligity makes accurate predictions for those cases without similar CDRH3s, we assessed the performance on both the "full set", and a subset where the CDRH3 identity is $\leq$0.8 ("CDRH3$\leq$0.8 set").

The number of positive and negative cases for these two sets are listed in Table 4.2. The performance at the selected thresholds is also reported as precision and recall (see Equations 2.4 and 2.5).

4.2.5 Selecting an epitope similarity threshold

To select an epitope similarity threshold, we carried out an evaluation based on the paratopes and epitopes of the co-crystal complexes. Paratopes in the crystal structures ("crystal paratopes") are defined as the antigen residues that have at least one atom within 4.5Å of the antigen; likewise for the "crystal epitopes". A set of hash tables and similarity scores were generated for the crystal paratopes and epitopes in the same way as for model paratopes (see Section 4.2.3).

We tested the classification performance by sweeping through pairs of crystal paratope and crystal epitope similarities in increments of 0.1, between 0 to 1. Consider a pair of antibodies with paratope similarity S_p and corresponding epitope similarity of S_e:

- True Positive (TP): $S_p >= S_p^t$ and $S_e >= S_e^t$
- True Negative (TN): $S_p < S_p^t$ and $S_e < S_e^t$
- False Positive (FP): $S_p >= S_p^t$ and $S_e < S_e^t$
- False Negative (FN): $S_p < S_p^t$ and $S_e >= S_e^t$.

We chose the epitope similarity score that gives the best classification performance, *i.e.* the highest Matthews correlation coefficient (MCC), that is calculated as follows:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}. \tag{4.2}$$

The selected epitope threshold becomes the ground truth for evaluation. For Ab-Ligity, we selected the crystal epitope similarity threshold at 0.1. Manual inspection of example cases indicated that pairs of epitopes above this threshold were highly similar. This threshold is used throughout the chapter to give the binary classification of similar and dissimilar epitopes.

4.2.6 Selecting a model paratope similarity threshold for real-life applications

We replicated a realistic scenario where only antibody models and predicted paratopes are available. To ensure the applicability of Ab-Ligity, we need to define a model paratope similarity threshold that can identify paratopes against highly similar crystal epitopes. Based on the definition of similar crystal epitopes found in the procedure above, we selected the model paratope similarity threshold by maximising the MCC.

4.2.7 Sensitivity analysis

We assessed the sensitivity of Ab-Ligity's performance in three situations: varying the distance bin size in hashing, changing the Parapred prediction threshold and applying Ab-Ligity to only the heavy (VH) or light chain (VL).

4.2.7.1 Distance bins

We tested the performance of Ab-Ligity when we used different bin sizes for the edge distance. For each distance bin size, we fixed the Ab-Ligity definition of similar epitopes to a score of 0.1, and found the corresponding Ab-Ligity score with the highest Matthews' correlation coefficient at the respective thresholds. For the bin size of 0.5Å, we kept the Ab-Ligity threshold for similar paratopes at 0.1, but we saw a slight increase in the Ab-Ligity threshold to 0.2 in the larger bin sizes.

4.2.7.2 Parapred thresholds

Since Ab-Ligity was developed to be used on predicted paratopes, we tested the effect of changing the Parapred threshold on the accuracies of Ab-Ligity and InterComp. The current Parapred threshold used in the manuscript is 0.67. We selected Parapred thresholds of 0.50 and 0.80 for this evaluation.

Reducing the Parapred threshold increased the number of residues in the CDR being predicted as the paratope, that is, the predicted paratopes became larger (Figure 4.2). This would also accentuate the noise by making more false positive predictions (Table 4.6). On the contrary, increasing the Parapred threshold reduced the predicted paratope sizes whilst increasing precision (Figure 4.2 and Table 4.6).

4.2.7.3 Predictions on heavy chains or light chains only

The majority of the currently-available immune repertoire datasets contains unpaired VH and VL sequences (Kovaltsuk et al., 2018). To assess the applicability of Ab-Ligity and InterComp on these datasets, we tested the classification performances of Ab-Ligity and InterComp on heavy chains or light chains only.

We carried out two tests. First we followed the original process outlined in the Section 4.2.2 to build paired homology models of antibodies using both the VH and VL sequences ("full antibody"). We then extracted the predicted paratope residues in each of the VH or VL chain, for the construction of paratope surfaces by Ab-Ligity and InterComp. This subset is referred to as "VH paratope" or "VL paratope".

The second test involved building homology models using a single VH or VL sequence, and taking the same set of paratope predictions as in the first test ("single domain antibody"). The distinction of the second test arises from the coordinates of the paratope residues: these coordinates could be different from the paired models as the "companion" chain was not present when calculating the structural clashes in the homology modelling process.

4.2.8 Benchmark

We compared our prediction performance and computational time with InterComp, a surface comparison tool for protein-protein interfaces (Mirabello and Wallner,

2018). We selected a qualitatively equivalent epitope similarity threshold (0.7) using the method outlined above, and evaluated the performance of both methods based on Ab-Ligity's and InterComp's definitions of similar crystal epitopes. The algorithm run-time was measured on a single 3.40GHz i7-6700 CPU core.

4.3 Results

Ab-Ligity compares two antibody paratopes by tokenized residues and distance hashes. It is designed to work on antibody models with predicted paratopes.

To test the power of Ab-Ligity to identify antibodies that bind to the same epitope, we simulated a real-life application of using antibody models and predicted paratopes. We built models of 920 unique protein-binding antibodies using ABodyBuilder (Leem et al., 2016), and predicted their paratope residues with Parapred (Liberis et al., 2018). Using these models and predicted paratopes we tested if Ab-Ligity was able to identify antibodies that bind to the same epitopes (as defined by the antibody-antigen crystal complexes; see Methods). We call this set the "full set". Since the majority of the similar binders have the same CDRH3 length and highly similar CDRH3 sequences, we also tested Ab-Ligity performance when we removed these "easy" comparisons (CDRH3 identity >0.8). We call this set the "CDRH3≤0.8 set". The number of comparisons for each of these tests are summarised in Table 4.2.

Table 4.2: Number of positive and negative comparisons in the datasets, based on Ab-Ligity's definition of similar epitopes.

Set	**Positive**	**Negative**
Full	719	29,557
CDRH3≤0.8	266	29,519

4.3.1 Selecting similarity thresholds

4.3.1.1 Definition of similar epitopes captured in crystal structures

In order to identify a standardised definition of when two epitopes are the same, we ran a grid search on the corresponding "crystal paratope" and "crystal epitope" similarity scores as calculated by Ab-Ligity, maximising the MCC (see Section 4.2.5). This procedure suggested an epitope similarity threshold of 0.1 was appropriate (see Appendix Table C.1) and visual inspections of examples confirmed this. Epitope pairs above this threshold are considered positive (*i.e.* similar; Table 4.2).

4.3.1.2 Selecting a similarity threshold for predicted paratopes in antibody models

In a real-life scenario where antibody models and predicted paratopes would be used, we need to establish a model paratope similarity threshold that can recapitulate the definition of similar epitopes as defined by the crystal structures. Similar to the earlier strategy, we ran a grid search on a range of "model paratope" similarity scores, and found a threshold of 0.1 that gives the best MCC (0.90, see Table 4.3). Above a model paratope similarity of 0.1, paratopes are predicted to bind to a common epitope.

Table 4.3: Performance of the selected thresholds based on model paratope and crystal epitope similarities defined by the same method.

Methods	Paratope Similarity	Epitope Similarity	MCC	Precision	Recall
Ab-Ligity	0.1	0.1	0.90	0.95	0.85
InterComp	0.6	0.7	0.81	0.80	0.83

4.3.2 Using Ab-Ligity to predict antibodies that bind to highly similar epitopes

We assessed the performance of Ab-Ligity on two datasets of different difficulties: the "easy" full set and the "hard" CDRH3≤0.8 set. Sequence-based methods can identify relatively accurately CDRH3 sequence-similar antibodies that will bind to the same epitope (Scheid et al., 2011). In our dataset, the majority of the antibody pairs that target highly similar epitopes have similar CDRH3 sequences. Table 4.2 shows that 63.0% (453/719) of the positive pairs that bind to highly similar epitopes have CDRH3 sequence identities >0.8. This set represents "easy" cases, and Ab-Ligity is able to stratify between similar and dissimilar binders with good accuracy on the "easy" full set (precision of 0.95 and recall of 0.85; Table 4.4).

To assess Ab-Ligity's performance on a more challenging set, we removed the "easy" comparisons. The remaining 37.0% (266/719; see Table 4.2) of the pairs that bind to similar epitopes have their CDRH3 sequence identities ≤0.8, and would not be identified by sequence-based methods such as clonotype. On this more challenging subset of cases, Ab-Ligity's precision remains at 0.95 and its recall drops slightly (0.69; Table 4.4). These results demonstrate that Ab-Ligity is able to accurately identify sequence-dissimilar antibodies that have similar binding modes.

Table 4.4: Precision and recall on the full and CDRH3≤0.8 sets, using the selected paratope similarity thresholds for Ab-Ligity (0.1) and InterComp (0.6), based on Ab-Ligity's definition of similar epitopes.

Method	**Ab-Ligity**		**InterComp**	
Set	**Precision**	**Recall**	**Precision**	**Recall**
Full	0.95	0.85	0.92	0.77
CDRH3≤0.8	0.95	0.69	0.86	0.59

4.3.3 Sensitivity analyses

We assessed the sensitivity of Ab-Ligity's performance in three situations: varying the distance bin size in hashing, changing the Parapred prediction threshold and applying Ab-Ligity to only the heavy (VH) or light chain (VL). The latter two factors change the paratope size. Using different Parapred thresholds affects the number of paratope residues predicted on each of the antibodies and potentially introduces noise if an inappropriate threshold is selected. We checked the performance on VH and VL alone as almost all current large sequencing datasets are unpaired (Kovaltsuk et al., 2018); if Ab-Ligity is able to accurately predict on a single chain, its potential application space is increased.

4.3.3.1 Distance bin size

During the hashing procedure, Ab-Ligity discretised the distances between residues (*i.e.* edge length of the triangles) into distance bins of 1Å. We observed that tightening the bin width to 0.5Å marginally reduces the classification performance with a precision and recall of 0.93 and 0.85 in the full set and a similar change on the CDRH3≤0.8 set (see Table 4.5). Increasing the bin size harms performance, potentially as it over-smooths residue distances (see Table 4.5).

Table 4.5: Performance of Ab-Ligity using different distance bin sizes on the two core sets, based on Ab-Ligity's definition of similar epitopes. P: Precision, R: Recall.

Distance bin size	**0.5Å**		**1.0Å(Original)**		**1.5Å**		**2.0Å**	
Ab-Ligity threshold	**0.1**		**0.1**		**0.2**		**0.2**	
	P	**R**	**P**	**R**	**P**	**R**	**P**	**R**
Full	0.93	0.85	0.95	0.85	0.98	0.74	0.82	0.56
CDRH3≤0.8	0.94	0.73	0.95	0.69	0.96	0.53	0.61	0.29

4.3.3.2 Parapred predictions

Ab-Ligity uses predicted paratopes, so we examined how their accuracy influences performance. The Parapred prediction threshold of 0.67 from the original paper was selected such that the size of the predicted paratopes replicates that of the actual paratopes. At this threshold, the precision and recall of Parapred were found to be 0.67 and 0.73 (Table 4.6). We increased the threshold to 0.80 (increasing precision and reducing the number of residues predicted; see Figure 4.2) to investigate the effect on Ab-Ligity. We saw little effect on the performance of Ab-Ligity: the precision and recall were 0.95 and 0.85 for the original Parapred threshold (0.67), compared to 0.96 and 0.80 in the increased Parapred thresholds (Table 4.7).

We then reduced the Parapred threshold, increasing the paratope size but lowering the precision of paratope prediction (Table 4.6). This reduction in the Parapred threshold decreased Ab-Ligity performance (Table 4.7). This drop in performance is probably due to an elevated level of noise (*i.e.* the inclusion of residues not involved in binding) in the Parapred predictions.

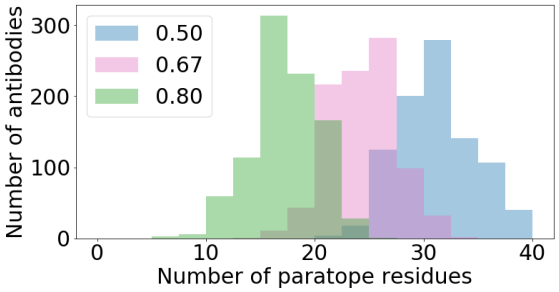


Figure 4.2: The size of the predicted paratopes in the non-redundant set, at different Parapred thresholds.

4.3.3.3 Performance on heavy chains or light chains alone

Since most publicly available immune repertoire datasets are from bulk sequencing of unpaired antibody chains (Kovaltsuk et al., 2018), we assessed the applicability

Table 4.6: Precision and recall of Parapred at selected thresholds.

	Parapred Performance	
Parapred Threshold	**Precision**	**Recall**
0.50	0.61	0.84
0.67 (Original)	0.67	0.73
0.80	0.73	0.55

Table 4.7: Performance of Ab-Ligity and InterComp with predicted paratopes extracted at different Parapred thresholds, on the full set. The paratope similarity scores for Ab-Ligity and InterComp were 0.1 and 0.6 respectively; while the epitope similarity scores were 0.1 and 0.7 for Ab-Ligity and InterComp.

		Epitope Definition			
Parapred Threshold	**Paratope Definition**	**Ab-Ligity**		**InterComp**	
		Precision	**Recall**	**Precision**	**Recall**
0.50	**Ab-Ligity**	0.73	0.25	0.66	0.28
	InterComp	0.87	0.23	0.79	0.25
0.67 (Original)	**Ab-Ligity**	0.95	0.85	0.83	0.92
	InterComp	0.92	0.77	0.80	0.84
0.80	**Ab-Ligity**	0.96	0.80	0.84	0.87
	InterComp	0.96	0.73	0.83	0.79

of Ab-Ligity on unpaired heavy and light chains.

We built homology models for heavy chain or light chain separately and extracted the predicted paratopes on these models. In Table 4.8, we show that heavy chain paratopes alone can be used to accurately identify antibodies that bind to the same epitopes, with a precision of 0.88 and recall of 0.78. Using light chain alone, Ab-Ligity can also identify antibodies that bind to similar epitopes but with lower precision (Table 4.8). We also evaluated using the heavy chain or light chain paratopes from paired antibody models and the results were almost identical (Table 4.8).

Table 4.8: Performance of Ab-Ligity and InterComp on heavy chain or light chain only paratopes, on the full set. Antibodies modelled using both VH and VL chains are labelled "full antibody", while those modelled using a single VH or VL chain only are labelled "single domain antibody".

			Epitope Definition			
Modelling	**Paratope regions**	**Paratope Definition**	**Ab-Ligity**		**InterComp**	
			Precision	**Recall**	**Precision**	**Recall**
Full antibody	**Original**	**Ab-Ligity**	0.95	0.85	0.83	0.92
		InterComp	0.92	0.77	0.80	0.83
	VH paratope	**Ab-Ligity**	0.90	0.78	0.78	0.84
		InterComp	0.75	0.80	0.65	0.85
	VL paratope	**Ab-Ligity**	0.64	0.90	0.54	0.94
		InterComp	0.16	0.93	0.13	0.95
Single domain antibody	**VH paratope**	**Ab-Ligity**	0.88	0.78	0.76	0.84
		InterComp	0.74	0.82	0.63	0.88
	VL paratope	**Ab-Ligity**	0.67	0.89	0.57	0.93
		InterComp	0.17	0.94	0.14	0.95

4.3.4 Comparing Ab-Ligity to InterComp

We compared Ab-Ligity to an existing general protein-protein interface comparison tool, InterComp. Since the original manuscript of InterComp did not indicate a threshold for interfaces to be considered similar, we conducted the same evaluation as for Ab-Ligity. The epitope similarity threshold was selected at 0.7 by maximising the MCC between crystal paratope and crystal epitope similarities (see Appendix Table C.1). The number of positive and negative cases stratified by the full and CDRH3 $\leq$0.8 sets shown in Table 4.9, gives consistent observation that 33.4% (193/578) of the antibodies with similar epitopes have CDRH3 sequence identities $\leq$0.8. Based on this definition of similar epitopes, 0.6 is the optimal cut-off for InterComp paratope similarity (Table 4.3).

Table 4.9: Number of positive and negative comparisons in the datasets, based on InterComp's definition of similar epitopes.

Set	**Positive**	**Negative**
Full	578	29,698
CDRH3$\leq$0.8	193	29,592

We assessed the performance of Ab-Ligity and InterComp using Ab-Ligity's or InterComp's definition of similar crystal epitopes. Tables 4.4 and 4.10 show that the two methods have comparable performance with Ab-Ligity showing better performance on the more challenging CDRH3≤0.8 set, regardless of whether the performance is assessed using the definition of ground truth by Ab-Ligity or InterComp (*i.e.* which pairs of antibody-antigen structures were considered to have similar epitopes by the two methods).

Table 4.10: Precision and recall on the full and CDRH3≤0.8 sets, using the selected paratope similarity thresholds for Ab-Ligity (0.1) and InterComp (0.6), based on InterComp's definition of similar epitopes.

Method	**Ab-Ligity**		**InterComp**	
Set	**Precision**	**Recall**	**Precision**	**Recall**
Full	0.83	0.92	0.80	0.83
CDRH3≤0.8	0.81	0.82	0.73	0.69

We carried out the same sensitivity analyses on InterComp as described above for Ab-Ligity. In terms of Parapred thresholds and predictions on VH or VL chains alone, very similar effects were observed (see Section 4.3.3).

As well as outperforming InterComp, Ab-Ligity is also significantly faster. We measured the algorithm run-time on a single 3.40GHz i7-6700 CPU core. On the full set of 920×920 pairwise comparisons, Ab-Ligity takes 0.5 CPU-minute to generate the 920 hash tables for all paratopes, and 18.5 minutes for an optimised all-against-all similarity calculation. InterComp does not require pre-processing but the all-against-all query takes 65.5 minutes. It is now possible to model large portions of next generation sequencing datasets (Raybould et al., 2020). Ab-Ligity would allow rapid comparison of binding sites in these datasets. For example, Ab-Ligity would take one day to process on 150 cores a next generation dataset of 100,000 antibodies, compared to five days for InterComp.

4.3.5 Anti-lysozyme antibodies with dissimilar CDRH3 sequences against highly similar epitopes

To show the power of Ab-Ligity to predict similar binding of antibodies with dissimilar CDRH3 sequences, we examined three anti-lysosyzme antibodies, HyHEL-26, HyHEL-10 L-Y50F mutant and HyHEL-63 (annotated as 1NDM_BA_C, 1J1O_HL_Y and 1NBY_BA_C respectively in Figure 4.3). These antibodies are known to all adopt the same binding mode against lysozyme (Figure 4.3A; Mohan et al., 2003). Their CDRH3 loops are seven residues long, but differ by at least three residues between any pair, so would not be considered to share the same clonotype (Figure 4.3B). We have also included a counter-example of the 3C8 antibody (annotated as 6OKM_HL_R in Figure 4.3) against Tumor Necrosis Factor Receptor superfamily member 4 (TNFRSF4). This antibody also has seven-residue long CDRH3 sequence and shares a maximum sequence identity of 0.57 with HyHEL-26 (1NDM_BA_C).

We used Ab-Ligity to calculate the similarity of their binding using models of predicted paratopes. As outlined above, an Ab-Ligity score of above 0.1 indicates that the antibodies in comparison have similar binding sites. The pairwise similarity scores of all three anti-lysozyme antibodies are all above 0.1, indicating that Ab-Ligity would classify them as targeting the same epitope. This classification is consistent with the observation in the co-crystal complexes (Figure 4.3A). Conversely, the Ab-Ligity scores of these three anti-lysozyme antibodies with the anti-TNFRSF4 antibody were all below 0.1. Reflected in the crystal structures shown in Figure 4.3A, the anti-TNFRSF4 antibody clearly binds to a different antigen and epitope to the other three antibodies.

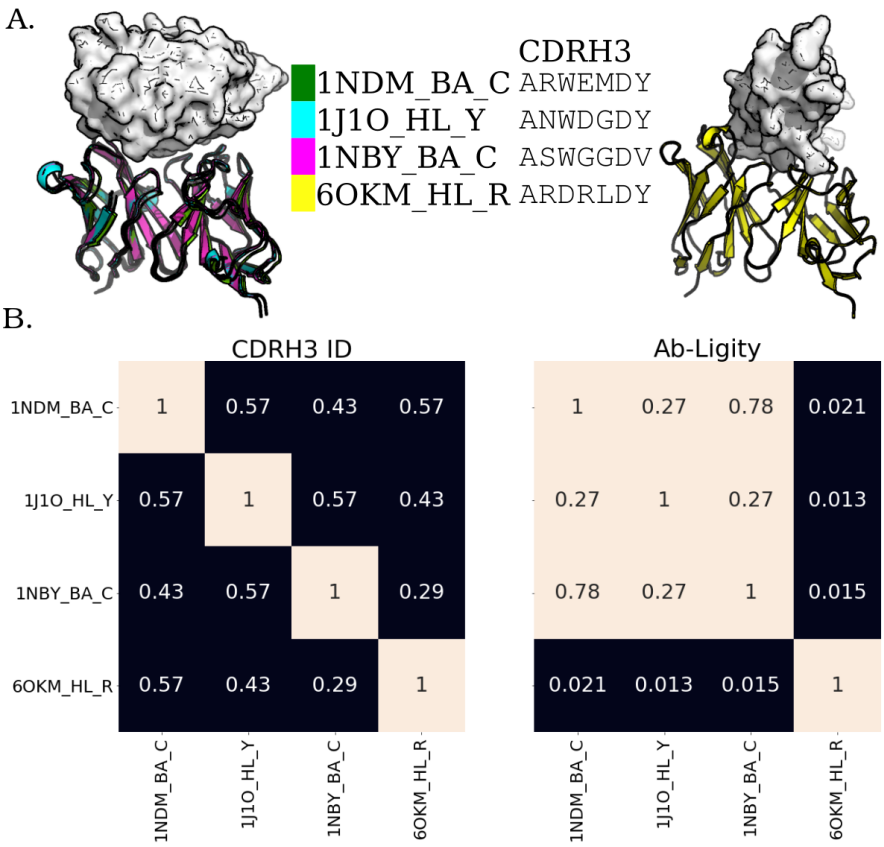


Figure 4.3: Analysis of anti-lysozyme antibodies with dissimilar CDRH3 sequences and highly similar epitopes. **A.** Structural superposition of three anti-lysozyme and one anti-TNFRSF4 antibodies co-crystallised with their antigens (lysozyme or TNFRSF4) in white. The three anti-lysozyme antibodies were HyHEL-26 (1NDM_BA_C), HyHEL-10 L-Y50F mutant (1J1O_HL_Y) and HyHEL-63 (1NBY_BA_C); and the anti-TNFRSF4 antibody is 3C8 (6OKM_HL_R). The antigens from the three anti-lysozyme antibody crystal structures are aligned. The legend shows the colours of the antibodies with their (PDB; Berman et al., 2000) codes followed by the heavy-light chain and antigen chain identifiers, separated by ("_"). The CDRH3 sequences are displayed next to the respective antibody identifiers. **B.** Heatmaps of CDRH3 sequence identity and Ab-Ligity paratope similarity. The row and column labels correspond to the structures shown in **A**. Ab-Ligity paratope similarity is calculated on the antibody model and predicted paratope as outlined in Section 4.2.2. Pairs of antibodies with CDRH3 identity of >0.80 would have been considered similar by sequence-based metric. For Ab-Ligity, a similarity score of >0.1 suggests that the antibodies bind to highly similar epitopes.

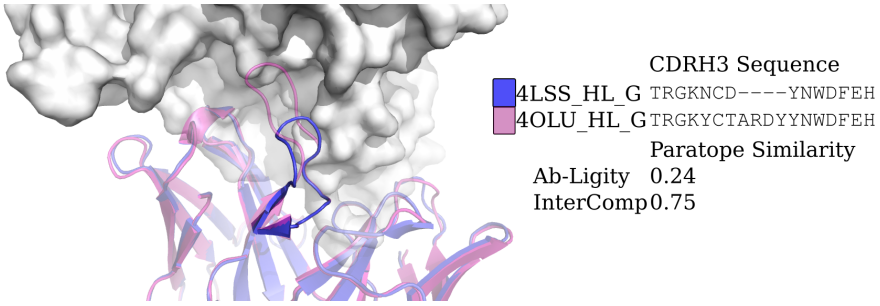


Figure 4.4: Analysis of two anti-gp120 antibodies. VRC01 antibody (PDB code and chain IDs: 4LSS_HL_G) is coloured in blue; VRC07 (PDB code and chain IDs: 4OLU_HL_G) is shown in pink. The gp120 core antigen is displayed as white surfaces and the superimposed antibodies are in cartoons. The CDRH3 loops are in solid shades of the cartoon representation. The PDB code, heavy and light chain ID, and antigen chain ID are separated by ("_") and listed in the legend with the corresponding CDRH3 sequences. CDRH3 sequences are shown by aligning their IMGT positions and "-" indicates a gap in the alignment according to the IMGT numbering scheme (Lefranc et al., 2015). The Ab-Ligity paratope similarity score for the pair is listed.

4.3.6 CDRH3 sequences with different lengths engage the same epitope in HIV core gp120

Ab-Ligity may also be useful in cases where antibodies have different CDRH3 lengths but similar binding modes. Sequence-based metrics are not applied to such cases (Galson et al., 2015). One example of mismatched CDRH3 lengths binding the same epitope are two HIV-1 neutralizing antibodies, VRC01 (PDB code and chain ID 4LSS_HL) and its variant, VRC07 (4OLU_HL). They target the gp120 core with highly similar binding modes (Figure 4.4). Their heavy and light chain germlines are the same but their CDRH3 sequences differ in length and composition. Aligning by the IMGT-numbered positions (Lefranc et al., 2015), only 12 out of the 18 residues are aligned and identical. These two antibodies are known to engage a similar epitope (Li et al., 2012) and Ab-Ligity based on models of these antibodies and predicted paratopes correctly predicts this (Ab-Ligity score of 0.24; $\geq$0.1 Ab-Ligity similarity threshold).

4.4 Discussion

In this chapter, we have presented Ab-Ligity, an antibody-protein binding site structural similarity metric that can identify sequence-dissimilar antibodies that engage the same epitope. Our results show that Ab-Ligity is able to identify antibodies that bind to the same epitopes using model structures and predicted paratopes, even for pairs of sequence-dissimilar antibodies. We evaluated the robustness of Ab-Ligity to the distance hashing, its dependence on the accuracy of paratope prediction by Parapred, and its applicability on unpaired antibody sequencing datasets. We also compared Ab-Ligity to InterComp, an existing protein surface similarity metric and found improved performance for harder cases with dissimilar sequences, and for the application on heavy chains or light chains only paratopes, and a far faster run-time. We further show that Ab-Ligity can identify antibodies that bind to the same epitope with dissimilar CDRH3 lengths, beyond the constraints of most sequence-based metrics.

As a structure-based metric, Ab-Ligity relies on homology modelling that is currently not scalable to full Ig-seq datasets. To apply Ab-Ligity on full Ig-seq datasets, other techniques can be used in conjunction. Apart from sequence-based clustering methods such as the clonotype analysis, a structural repertoire analysis (Kovaltsuk et al., 2020; Raybould et al., 2020; Marks and Deane, 2020) can extract the structurally-unique sets of sequences in repertoires. Since this clustering implies that multiple sequences may map to the same framework and CDR backbone structures, Ab-Ligity may first need to be adapted such that it downplays the importance of the exact residue identity but focuses on the spatial (distance) information of the paratope. Alternatively, based on a set of public antibody structures commonly found across multiple patients' immune repertoire (Raybould et al., 2020), it may also be possible to pre-build a database of Ab-Ligity hash tables. The database of representative structures can then be screened against antibodies of interest to find any potential structurally-similar hits in the repertoire datasets.

Binding site comparison with tools, such as Ab-Ligity, opens up an alternative way to search for binders with similar binding modes. Typically in antibody discovery, multiple diverse hits are desired to avoid developability issues in the downstream optimisation process (Raybould et al., 2019a). Ab-Ligity has the ability to predict if antibodies share similar target epitopes, without being sequence-similar. This ability coupled with the fast run-time of Ab-Ligity makes it suitable for searching through large datasets of antibody sequences to find sets of sequence-diverse binders to the same epitope.

4.5 Chapter Summary

We described Ab-Ligity, a novel hashing method of describing paratopes and comparing their similarity to find paratopes against the same epitope. We tested its sensitivity on the hashing parameters and found that the algorithm is reasonably robust to variations in the distance bin sizes, paratope predictions and interface sizes. We compared Ab-Ligity with an existing protein surface comparison tool, InterComp. Ab-Ligity has an improved performance on the more difficult cases with distinct sequences. As most of the publicly available antibody sequences are unpaired, we compared the applicability of both tools on this set and found that Ab-Ligity can handle heavy chains or light chains only paratopes better than InterComp. We further showed that Ab-Ligity can find sequence-dissimilar but structurally-similar paratopes against the same epitopes, demonstrating its ability to move beyond sequence-based metrics.

In the next chapter, we will further study paratope features. We will evaluate an existing paratope prediction tool, Parapred (Liberis et al., 2018), and assess its behaviour stratified by CDR loop lengths. We will also present some preliminary results on the important paratope features and how they can be used to build models for paratope prediction.

5
Paratope analysis

Contents

5.1 Introduction

The previous chapter described Ab-Ligity, a method that compares binding site structures and identifies sequence-dissimilar antibodies that bind to the same epitope. Since Ab-Ligity uses paratopes predicted by Parapred (Liberis et al., 2018), we assessed Ab-Ligity's sensitivity over different Parapred prediction thresholds and confirmed its robustness on a set of public antibody structures. In this chapter, we further study the behaviour of Parapred and investigate what features could be used to predict paratopes engaging different types of antigens.

The structural dataset available only covers a limited subset of the totality of paratope space. On of the key factors of specificity – CDRH3 lengths (Barrios et al.,

2004), Kovaltsuk et al. (2020) found that the natural repertoire and structural data have different preferences. The CDRH3 length distributions in human and mouse repertoires peak around 14-15 and 11-12 respectively, while that of the overall (species-agnostic) structural data is around 12. Only half of the CDRH3 in the human repertoire datasets used in Kovaltsuk et al. (2020) could be confidently annotated with the currently available database-search structural modelling tool. This limitation in the applicability on repertoire data, may be expected in other tools trained on structural dataset alone, such as sequence-based paratope prediction.

Sequence-based paratope prediction tools mainly restrict their predictions to around the CDRs and use physicochemical properties as one of the input features to represent the residues. To infer the relationship between residues and their positions in a sequence, Parapred use a recurrent neural network layer (Liberis et al., 2018), while proABC used a random forest (Olimpieri et al., 2013) and in proABC-2, a convolutional neural network (Ambrosetti et al., 2020). One of the earlier tools, Paratome, ignored the physicochemical properties (Kunik et al., 2012a). Instead, relying solely on the residue positions, Paratome found the closest sequence match (or a homolog) to the query sequence, and transferred the paratope positions on the homolog to the query antibody. It suggests that residue position is likely to be a deterministic factor for paratope prediction, and this rule-based model is sufficient to give a minimal benchmark for all forthcoming machine-learning models.

Different CDR types have varying propensity to engage the antigen (Nguyen et al., 2017). CDRH3 is the most variable in sequence and structure, and is almost always in contact with the antigen. On the other hand, CDRL2 has only a couple of highly conserved canonical forms and is less often found in the proximity of the antigen. Different CDR types may use different binding motifs and should be considered independently.

Antibodies can engage antigens of different types and sizes, but most paratope predictors focus on protein-binding paratopes only (see Section 1.4.2.1). To complete our understanding of antibody-antigen binding, we assessed the applicability of such frameworks on other types of antigens. Since the paratope shapes were found to vary according to the size of the antigens (see Section 1.4.2.2), the binding patterns are likely to be different. Training paratope predictors against specific antigen types and sizes could potentially inform the paratope features unique to different types of antigens leading to more accurate predictions.

One of the key outputs of computational models is to identify and explain hidden patterns (*e.g.* Olimpieri et al., 2013). Deep learning frameworks such as Parapred and proABC-2 face a challenge with the interpretability of their models. Due to the model complexity, it is difficult to decipher which features have contributed to the predictions and how the features correlate with each other. Feature importance in neural networks is usually assessed through ablation studies, where each of the features is left out one-by-one to train the model and the drop in accuracy reflects the importance of that feature (*e.g.* Gainza et al., 2020). This can be a time-consuming process to refit the model for every single, or an ensemble, of features. On the other hand, traditional statistical models are less complex. In linear models, the coefficients for each feature can be used directly to indicate the positive or negative contribution of the feature to the overall prediction. For ensemble or tree-based models, two main types of metrics have been compared: mean decrease in impurity (MDI) and permutation importance (Strobl et al., 2007; Altmann et al., 2010). In random forest, each node in the decision trees determine how to split based on a locally optimal condition known as "impurity". In MDI, the decrease in the Gini impurity (or information entropy) is assessed for each feature (Strobl et al., 2007). On the other hand, permutation importance measures the model score when a single feature value is randomly shuffled (Altmann et al., 2010). It has been shown that MDI tended to upweight continuous variables over categorical variables, while permutation importance less so (Strobl et al., 2007). Depending on the features'

data types, the two metrics may not necessarily return the same importance ranking. However, computing these metrics is still more efficient than the ablation studies in deep learning models.

In this chapter, I describe a series of initial studies into how to improve paratope prediction. We studied the behaviour of Parapred under different scenarios, and attempted to construct more transparent statistical models. First, we examined the positional dependence of the real and Parapred-predicted paratopes. We then studied if binding propensity was affected by CDR structural clusters (*i.e.* canonical forms). For the residue features, we used two types of physicochemical feature representation schemes: categorical and continuous values. To consider the neighbourhood contribution, we included the features from the surrounding residues. We assessed if a simple model could capture these features. LASSO and random forest were used, and their performances and behaviours were compared with Parapred. Specifically, we investigated the positional dependence of the predictions, and their ability to generalise to longer CDR sequences when trained only on shorter loops. To assess if these frameworks could be transferrable to other antigen types and find if the paratope features were similar between protein-binding and peptide-binding antibodies, we retrained the models on peptide-binding antibodies and compared the feature importance with that obtained from the original protein-binding set.

5.2 Method

5.2.1 Datasets

We ran the preliminary analysis using the original set of antibodies in the Parapred paper (Liberis et al., 2018), but with the numbering scheme and CDR definition as defined by the IMGT scheme (Lefranc et al., 2015). In short, this set of 277 non-redundant antibodies all bound to protein antigens and had been filtered at 95% sequence identity. Their six CDRs, and two framework residues immediately

before and two after each CDR, were extracted. In Parapred, loops were aggregated regardless of their CDR types for training, *i.e.* 1662 CDR sequences were used in the Parapred training set, hereafter the "Parapred set". For the analysis in this chapter and our paratope prediction models, loops were separated by their CDR types. For example, in the CDRH1 analysis and the subsequent predictor, 277 sequences were used.

5.2.2 Features

We assessed three features to describe the residues: by IMGT position, canonical form or sequence length and their physicochemical features (see Figure 5.1). Features are one-hot-encoded unless they are continuous values (*i.e.* Meiler's features, see below). We also examined if the neighbourhood was important by applying different sequence padding.

5.2.2.1 Position

We recorded the IMGT positions of each residue. For example, in a CDRH1 sequence with two anchor residues each on both termini `VSGFTFTDFYMN`, the third residue `G` (Glycine) will have its "position" feature noted as 27 because it is the IMGT position of the start of CDRH1.

5.2.2.2 Canonical forms or sequence length bins

To annotate the canonical forms of the CDR sequences, we used SCALOP (described in Chapter 2), a sequence-based canonical form prediction tool for non-CDRH3 loops, with IMGT numbering and CDR definition. For CDRH3, we classified the loops by their length (minus the four additional residues): <13, 13–16 and >16. This was taken from the analysis in SAAB+ (Kovaltsuk et al., 2020) where sequences in these bins have the same score cut-off to indicate similar confidence level in

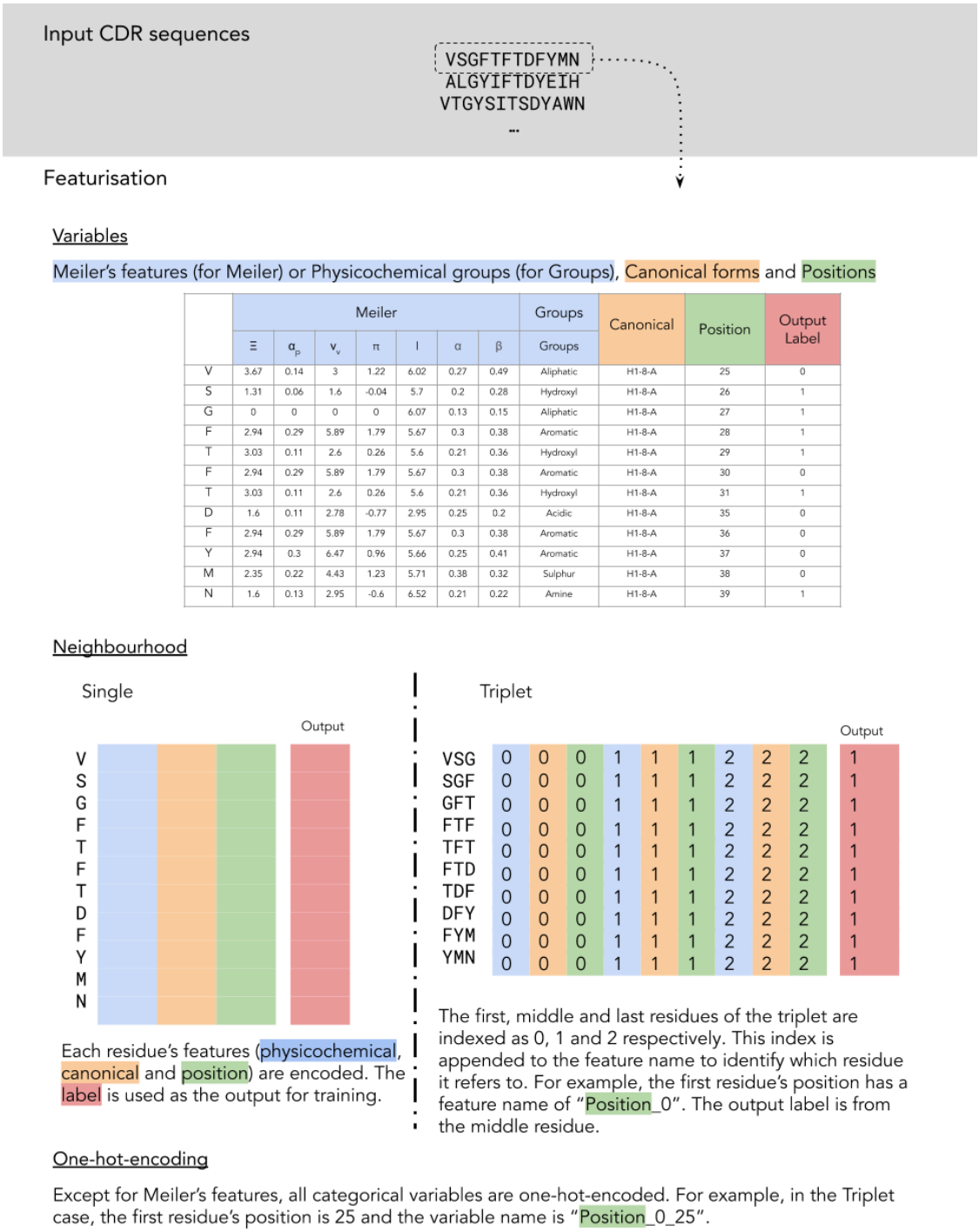


Figure 5.1: Paratope prediction feature encoding pipeline. This example is from CDRH1. The set of input CDRH1 sequences (and two framework residues each on both termini) were extracted. The physicochemical features – "Groups" or "Meiler", the canonical forms and the residue positions are recorded as the input feature matrix, and the corresponding label of whether the residue is in the paratope (1 for paratope, 0 for non-paratope) is used as the output. The neighbourhood encoding divides into two types: Single and Triplet. In Single, the residues are encoded on their own. In Triplet, trimers of residues are encoded using only the label of the middle residue as the output. Finally, one-hot-encoding is applied to all features except the continuous Meiler's features.

modelling.

5.2.2.3 Physicochemical features

We encoded the physicochemical properties of residues in two ways: with continuous values as in the Meiler's features (Meiler et al., 2001), or with categorical labels as in residue groups (Figure 1.4, Table 4.1).

Meiler's features

In Parapred, seven physicochemical properties were used to encode the residues (Table 5.1; Meiler et al., 2001). These features may affect the interactions between two proteins and the values quantify their level of effects. In short:

- Steric parameter (Ξ) is the graph shape index that is based on the complexity, branching and symmetry of the side chain;
- Hydrophobicity (π) is the $log(P_{amino-acid}) - log(P_{glycine})$ where P is the partition coefficient of the amino acid in octanol/water;
- van der Waals volume (ν_v) is the side chain volume excluding the volume of the hydrogen, normalised by the volume of linear aliphatic chain (-CH_2), using the equation $\nu_v = \frac{V_{side-chain} - V_H}{V_{CH_2}}$where V is the volume;
- Polarisability (α_p) considers the molecular weight, density and the index of refraction;
- α-helix propensity (α); and
- β-sheet propensity (β).

Residue groups

As described in Figure 1.4, Table 4.1 and used in Chapter 4, we tokenised the paratope residues by their physicochemical properties.

Table 5.1: Meiler's features. Ξ: steric parameter, α_p: polarisability, ν_v: volume, π: hydrophobicity, **I**: isoelectric point, α: helix probability, β: sheet probability.

Amino Acid	Ξ	α_p	ν_v	π	**I**	α	β
C	1.77	0.13	2.43	1.54	6.35	0.17	0.41
S	1.31	0.06	1.6	-0.04	5.7	0.2	0.28
T	3.03	0.11	2.6	0.26	5.6	0.21	0.36
P	2.67	0	2.72	0.72	6.8	0.13	0.34
A	1.28	0.05	1	0.31	6.11	0.42	0.23
G	0	0	0	0	6.07	0.13	0.15
N	1.6	0.13	2.95	-0.6	6.52	0.21	0.22
D	1.6	0.11	2.78	-0.77	2.95	0.25	0.2
E	1.56	0.15	3.78	-0.64	3.09	0.42	0.21
Q	1.56	0.18	3.95	-0.22	5.65	0.36	0.25
H	2.99	0.23	4.66	0.13	7.69	0.27	0.3
R	2.34	0.29	6.13	-1.01	10.74	0.36	0.25
K	1.89	0.22	4.77	-0.99	9.99	0.32	0.27
M	2.35	0.22	4.43	1.23	5.71	0.38	0.32
I	4.19	0.19	4	1.8	6.04	0.3	0.45
L	2.59	0.19	4	1.7	6.04	0.39	0.31
V	3.67	0.14	3	1.22	6.02	0.27	0.49
F	2.94	0.29	5.89	1.79	5.67	0.3	0.38
Y	2.94	0.3	6.47	0.96	5.66	0.25	0.41
W	3.21	0.41	8.08	2.25	5.94	0.32	0.42

5.2.2.4 Neighbourhood

We assessed if encoding the sequence into sliding windows of three (considering one residue before and one after the "centre" residue) would contribute to paratope characterisation. This is known as the "Triplet" featurisation (see Section 5.2.3.1 for more details on nomenclature). Each CDR sequence was broken down into a set of triplets. For example, the CDRH1 sequence `VSGFTFTDFYMN` was broken down into ten triplets: `VSG`, `SGF` and so on. The centre residue's label (of whether it is an actual paratope; see Section 5.2.3) was used for the triplet as the output, for the model training and testing. Each of the residue's features were individually identified to indicate if the feature belonged to the first, second or third residue in the triplet.

We also considered single-residue prediction. In this case, the starting and ending residues at the two ends of a CDR loop were also considered in the performance evaluation, and plotted in the positional frequency evaluation (see Section 5.2.4).

In Triplet, no predictions were made on the first and last residues of a sequence. In the calculation of recall and precision, the first and last residues were not considered. However for visualisation purpose on the positional frequency plots (see Section 5.2.4), the first and last residues were padded in with 0 and indicated as a lack of prediction.

5.2.3 Models

We used the original framework of Parapred (Liberis et al., 2018). Briefly, they padded the sequences to ensure equal lengths. This input matrix was then processed by a convolutional neural network layer, followed by a type of recurrent neural network model – the bidirectional Long Short-Term Memory (LSTM). Finally they used a drop-out layer to prevent overfitting.

In our work, we used two simple statistical models, a linear LASSO model and a tree-based random forest (RF) model, implemented in *scikit-learn* (Pedregosa et al., 2011). We used the Parapred set separated by CDR types, as input to our models. The output labels were set as 1 for paratope, and 0 for non-paratope (a classification task). Thus, LASSO was implemented using `LogisticRegression` with L1-norm as the penalty term and $max\_iter = 100$, while RF used the function `RandomForestClassifier` with the `n_estimators=100` (*i.e.* 100 trees).

To extract the feature importance in LASSO, we took the absolute values of the feature coefficients. For RF, we used the mean permutation importance of the features from the five-fold cross-validation run (see Section 5.2.4). As we have a mixture of continuous and categorical variables, using the default MDI may skew

the importance in favour for the continuous variables.

5.2.3.1 Nomenclature of the models

We named the combination as follows: the model name, followed by the neighbourhood and then the physicochemical featurisation. The model name is either LASSO, or "RF" for random forest. The neighbourhood (see Section 5.2.2.4) is noted as "Triplet" for the trimers, and "Single" for considering only a single residue. Physicochemical featurisation as noted in Section 5.2.2.3 can be done in two ways: "Groups" for the seven categorical residue groups and "Meiler" for the continuous values reported in Meiler et al. (2001).

5.2.4 Performance evaluation

5.2.4.1 Performance metrics

To evaluate the performance of our predictors, we used two metrics. First, we reported the recall and precision (Equations 2.4 and 2.5) of the models at the selected thresholds. In LASSO models, we selected the CDR-specific thresholds for prediction by maximising the Matthews correlation coefficient (MCC; Equation 4.2). For random forest models, we used 0.5 as the threshold. Second, to assess the positional dependence of the models, we plotted the relative positional frequencies of the predictions. For each IMGT position, we counted the frequency of the position being an actual paratope and a predicted paratope. To calculate the relative frequency, we normalised the frequency by the occurrence of the position of interest. For example, if position 112A only appears 60 times out of all 100 sequences, and 59 times it was actually a paratope but was only predicted as one by Parapred 48 times, the relative frequency of "actual paratope" would be 59/60 and that of the "predicted paratope" would be 48/60. This does not capture if the predictions were correct, *i.e.* the frequency includes all correct and incorrect predictions made at that IMGT position. For Parapred, we used the threshold of

0.67 as suggested in the original paper (Liberis et al., 2018). For our models, we used the thresholds obtained as described above.

5.2.4.2 Evaluation settings

The evaluation was conducted in three settings: cross-validation, blind set and length-dependent stratification.

Parapred used ten-fold cross-validation, run over 10 different random seeds. For our predictors, due to the smaller training set (277 sequences per CDR type), we carried out five-fold cross-validation.

In the "blind set" evaluation, predictors were trained on the entire Parapred set, with Parapred still aggregating all CDR types and our models separated by CDR types. A blind set was curated from antibodies that were unseen in the Parapred set and released before 3rd February, 2020. The same quality criteria was applied to retain only paired antibodies that have $\leq$3Å resolution, are non-protein binding, and are $\leq$95% full-length sequence-identical to the antibodies in the Parapred set and within the blind set. In total, there were 328 antibodies in this set. Within this set, we retained the CDR sequences unseen in the Parapred set and unique within the blind set, leaving 250 H1, 286 H2, 319 H3, 223 L1, 117 L2 and 289 L3 sequences. These are listed in Appendix Table D.1.

We assessed the length-dependence of the predictors by using CDR loops with $<$14 residues (excluding the four anchor residues) as the training set and those with $\geq$14 residues as test set. We retrained Parapred with this new "length-stratified set", still aggregating different CDR types as in the original paper. For our predictors, we separated the training by CDR types. Only three of the CDR types had loops $\geq$14 residues long: 2 in CDRH1, 126 in CDRH3 and 1 in CDRL3. We only inspected

CDRH3 in our study as the dataset sizes for the other two CDR types were too small.

5.2.5 Peptide-binding antibodies

To assess if the predictor models may be applicable to describe paratopes against other types of antigens, we curated a set of peptide-binding antibodies. We used the same quality cut-off and redundancy threshold as outlined in the Parapred paper, from all the antibodies released before 16$^{\text{th}}$ June, 2017 (the latest release date in the Parapred set). This resulted in 184 antibodies, and 1104 CDR sequences (six CDRs from each antibodies), collectively known as the "peptide set" for the aggregated training for Parapred. The peptide set was separated by CDR types to train our LASSO and RF models. These PDB codes are listed in Appendix Table D.2.

We retrained Parapred using the same architecture and hyperparameters, aggregating all CDR types. For our predictors, we used the same model architecture but retrained by CDR types as described above. The precision, recall and positional frequencies of the predictions were reported to assess their performance. The feature importance in the LASSO and RF models was extracted in the same way as above.

5.2.6 Feature importance analysis

To compare the feature importance in protein-binding and peptide-binding antibodies, we separated the analysis by CDR types. We extracted the feature importance scores from the LASSO and RF models, from the five-fold cross-validation runs. We then ranked the average importance scores of each feature, in descending order, *i.e.* the feature with the highest average importance score had the lowest rank index. Where two features had the same score, both took the lowest rank index (down to an integer number). We then calculated the Pearson correlation coefficient between the features between the rank number, that is equivalent to the Spearman's rank correlation coefficient (ρ). A Spearman's rank correlation coefficient of 1 means the

two variables are highly correlated, and -1 means they are inversely correlated. A ρ of >0.7 is generally considered a good correlation.

5.3 Results

We evaluated the positional preference of Parapred predictions, and found that the middle section of long loops was rarely predicted by Parapred. We then proposed features that potentially contribute to paratope prediction, including CDR type, canonical forms/sequence lengths, physicochemical properties and the neighbourhood. We built statistical models using LASSO and random forest to test if the simpler frameworks could sufficiently capture these paratope features while relaxing the length dependence seen in Parapred. To simulate the case of training paratope prediction tools on shorter loops in crystal structures but applying them on longer loops in repertoire datasets, we stratified the training set by sequence lengths to test the behaviour of Parapred and the simpler statistical models. Finally, we contrasted the features used by the interpretable LASSO and random forest models to predict paratopes in protein-binding and peptide-binding antibodies.

5.3.1 Positional frequencies of actual and Parapred-predicted paratopes

We first examined the positional preference of real and Parapred-predicted paratopes. Figure 5.2 shows that on CDRH2 and L3, Parapred made predictions at slightly higher frequencies than in the actual paratope, but on the same set of preferred positions. In both CDRH1 and L1, Parapred did not make any predictions at the start of the loop, even though in ∼20% of the time, they were in the vicinity of the antigen. The opposite occurred in L2 where the middle of the loop was falsely predicted as paratopes even though these positions were never in contact in real structures. CDRH3 has the most variable sequence length, ranging from six to over thirty. Only the longer loops have positions between 111C and 112C. The

relative frequency plot of CDRH3 suggests that when the positions 111C – 112C exist, these positions are almost always in the true paratope. However, Parapred rarely predicted these residues to be part of the paratope.

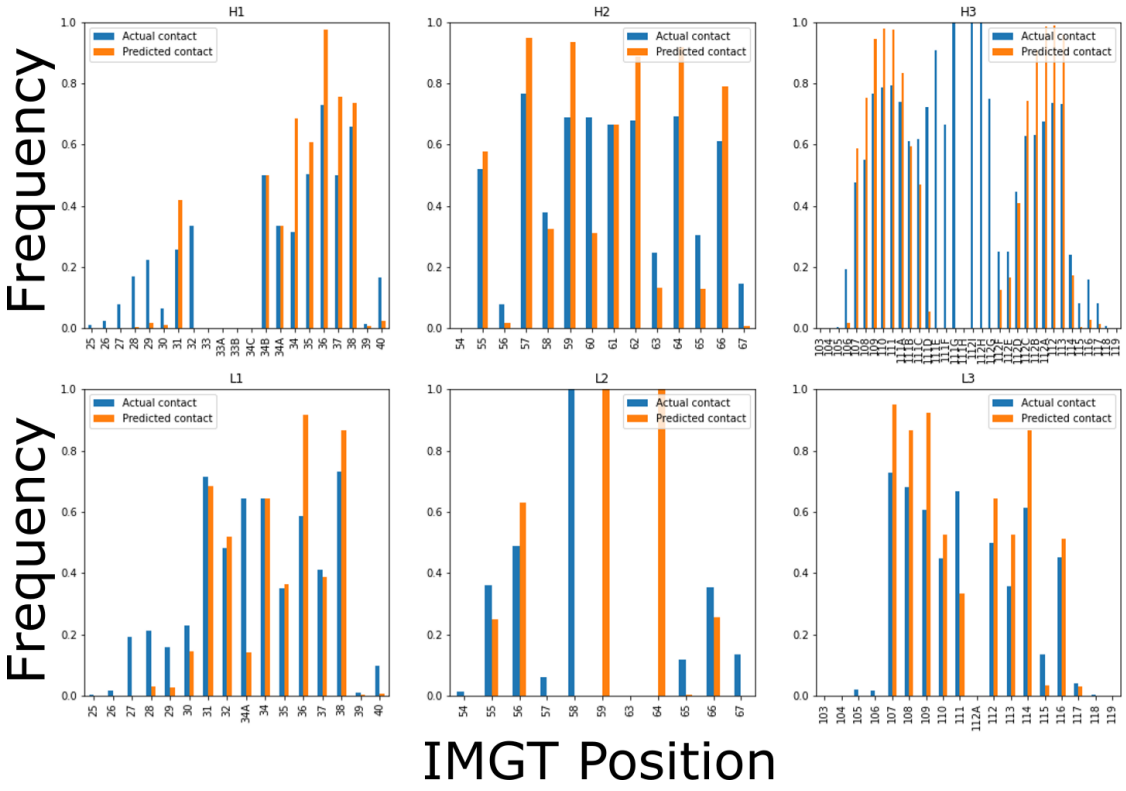


Figure 5.2: Relative frequencies of the actual and Parapred-predicted paratopes by IMGT-positions for each CDR types, in the Parapred set. Blue bars represent the relative frequencies of the actual paratope, and the orange bars represent those of the Parapred-predicted paratopes.

We then tested if the positional frequencies of the actual paratopes varied between the CDR canonical forms in non-H3 CDRs, or among different sequence length bins for CDRH3. We found that between different canonical forms, the actual positional binding probability varied (see Figures 5.3 – 5.8). In CDRH1 (Figure 5.3), residues in positions 27 – 29 occasionally participated in binding when they were found in loops shorter than 10 residues, in stark contrast to the length-10 canonical forms. Figure 5.4 suggests that positional preferences in CDRH2 are different in different canonical forms. Positions 58, 59 and 64 of CDRH2 showed different

binding patterns among the canonical forms. The frequency plots separated by CDRH3 lengths (Figure 5.5) show that that regardless of the sequence lengths, residues in CDRH3 were very often involved in binding, consistent with the general observation that the highly variable CDRH3 engages with the antigen (Nguyen et al., 2017).

CDRL1 has more canonical forms than any CDR type (Figure 5.6). Position 38 in CDRL1 was often found in the binding site regardless of the canonical forms of CDRL1. In most CDRL1 canonical forms, the frequency peaked at the middle of the loop, except in L1-8-A. In CDRL2, only one canonical form was observed. The mid-section of the three-residue-long L2-3-A canonical form was not often involved in binding (Figure 5.7). In CDRL3 (Figure 5.8), L3-5-A regularly used the positions 107, 116 and 117 for binding, while the other L3 canonical forms exhibited different usage patterns.

5.3.2 Cross-validation performance

Based on the observations above, we hypothesised that residue position and canonical form play an important role in paratope identification. We further considered the neighbourhood and contrasted the use of categorical or continuous physicochemical property representations of the residues as input for our simpler paratope prediction models. As outlined in Section 5.2.4, we assessed the performance of the LASSO and random forest (RF) models by cross-validation and selected an optimal threshold to balance between their precision and recall (see Table 5.2). Figure 5.9 shows that Parapred was better than any of our combinations in most cases. Our best-performing model was the RF Triplet Meiler model, yet its performance was still behind that of Parapred by 0.05 – 0.1 in precision and recall.

We also evaluated the frequencies of the predictions by IMGT-positions (Figures 5.10 – 5.11). The different prediction models tested did in some cases improve

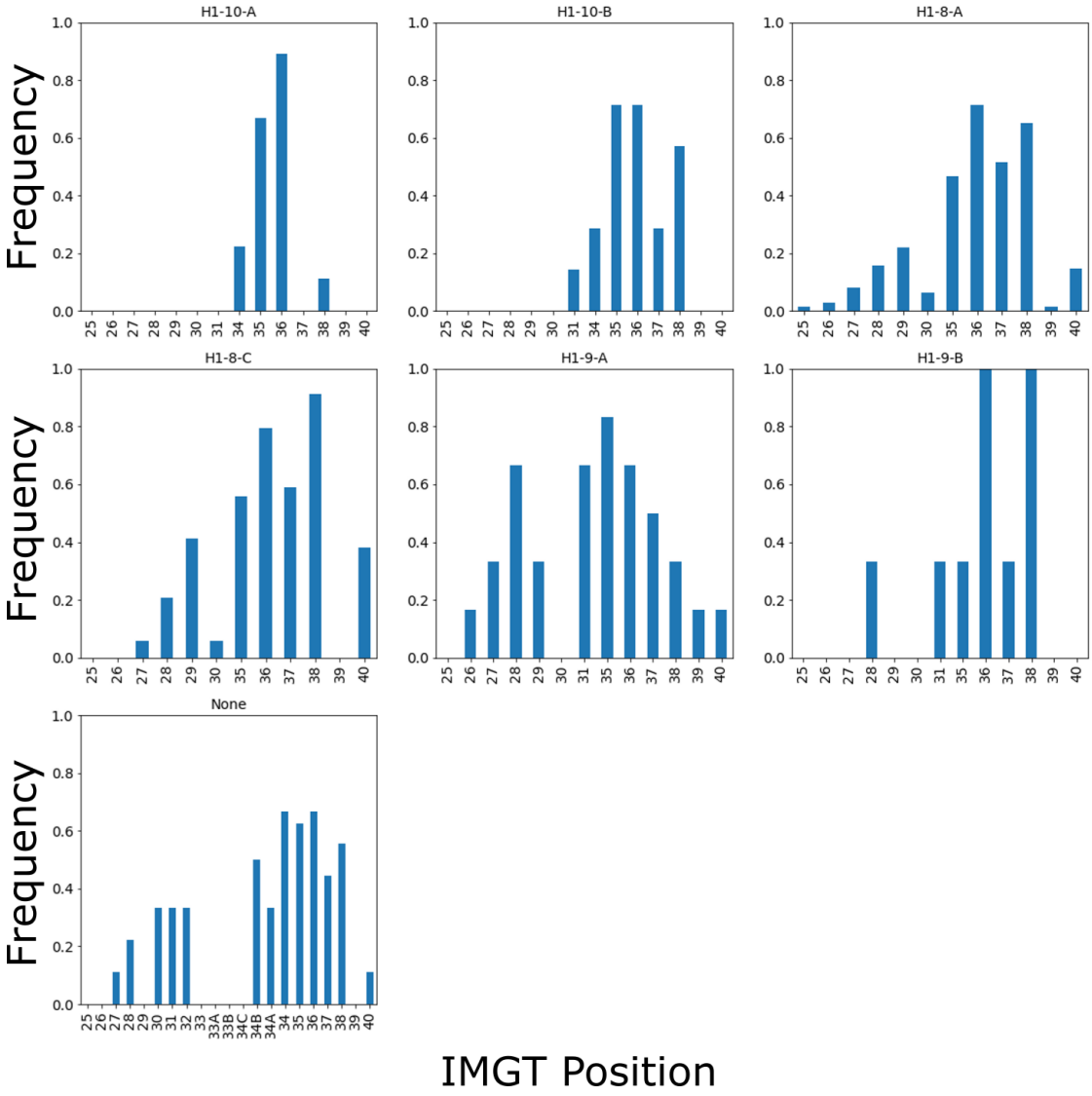


Figure 5.3: Relative frequencies of the actual paratopes by IMGT-positions, stratified by CDRH1 canonical forms in the Parapred set. IMGT-defined CDRs were used. Two residues before and two after the CDRs were also plotted.

positional prediction over Parapred, but not enough to improve overall accuracy.

5.3.3 Blind set

We tested the performance of these predictors on an unseen set of sequences. Parapred still retained the best performance out of all the models (Figure 5.12). Our models achieved similar precision and recall amongst themselves. When

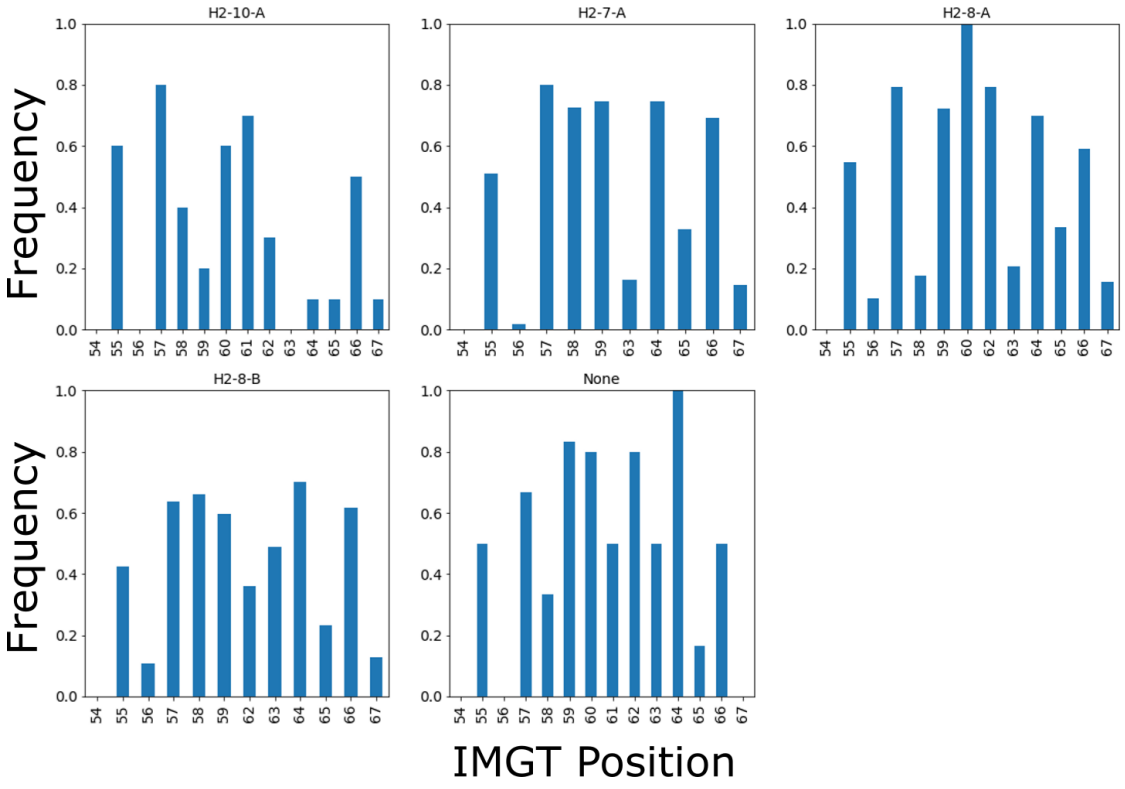


Figure 5.4: Relative frequencies of the actual paratopes by IMGT-positions, stratified by CDRH2 canonical forms in the Parapred set. See Figure 5.3 for details.

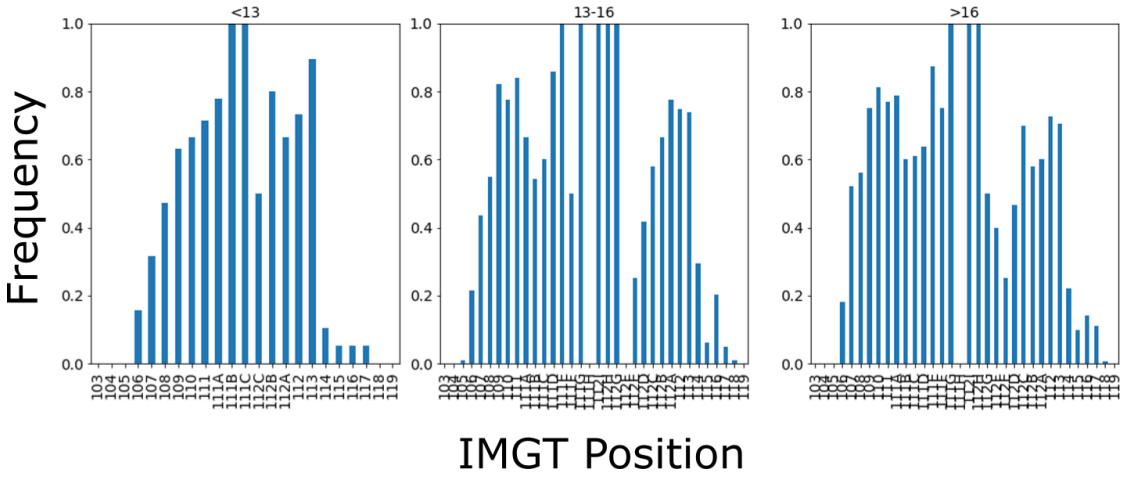


Figure 5.5: Relative frequencies of the actual paratopes by IMGT-positions, stratified by CDRH3 sequence lengths in the Parapred set. See Figure 5.3 for details.

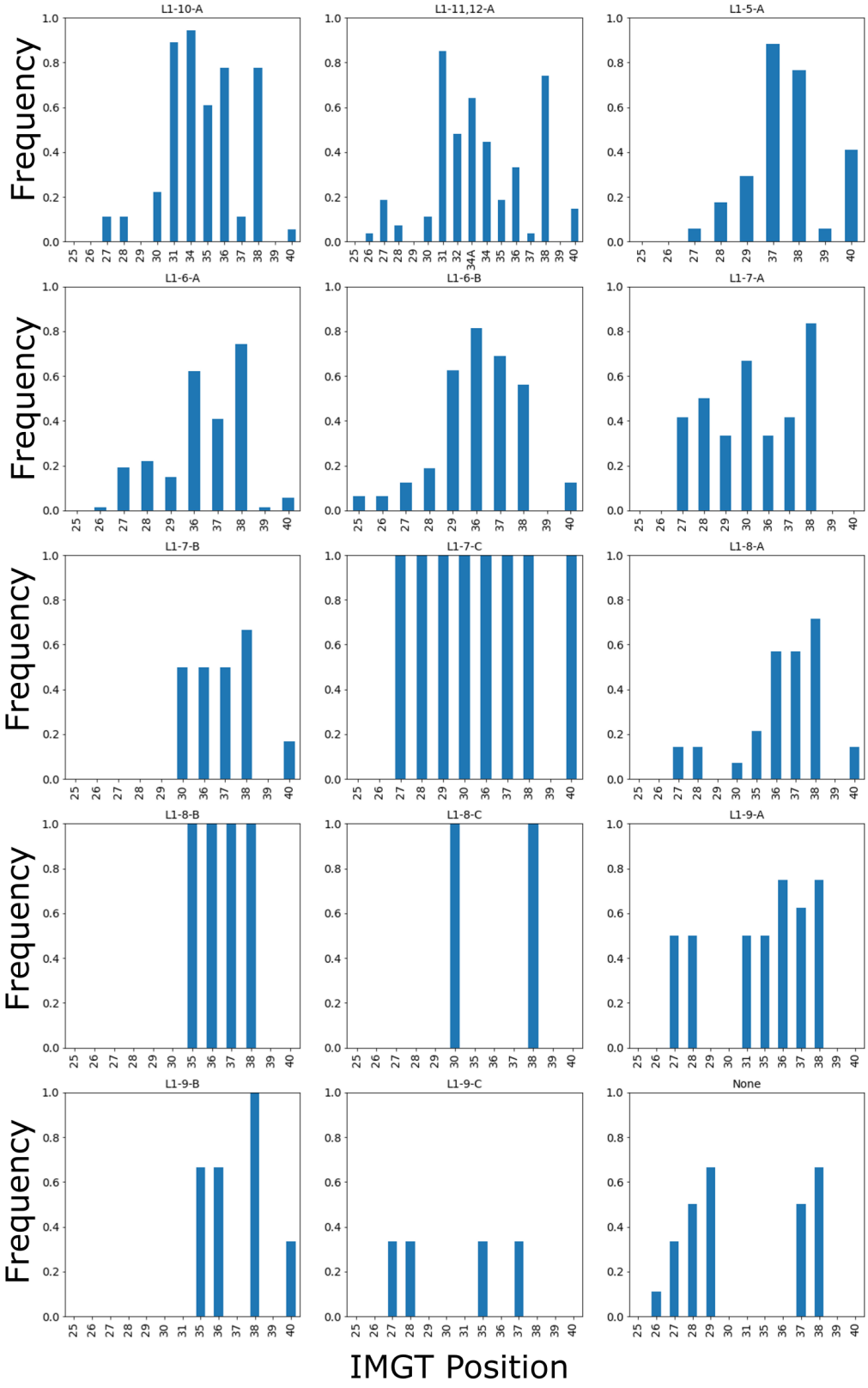


Figure 5.6: Relative frequencies of the actual paratopes by IMGT-positions, stratified by CDRL1 canonical forms in the Parapred set. See Figure 5.3 for details.

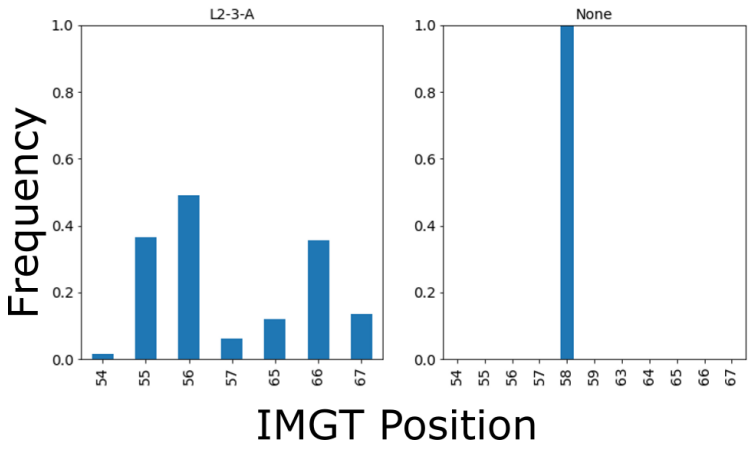


Figure 5.7: Relative frequencies of the actual paratopes by IMGT-positions, stratified by CDRL2 canonical forms in the Parapred set. See Figure 5.3 for details.

Table 5.2: Selected thresholds for the Parapred set. Refer to Section 5.2.3.1 for the model nomenclature. In short, "LASSO" and "RF" refer to the linear model and random forest model. "Single" and "Triplet" specify whether the sequences are broken up by residue or in a sliding window of three residues. "Groups" refers to the physicochemical tokens (Figure 1.4, Table 4.1), while "Meiler" is encoded by the Meiler's feature (Meiler et al., 2001).

Featurisation	**Triplet Groups**		**Single Meiler**		**Triplet Meiler**		**Parapred**
Model	**LASSO**	**RF**	**LASSO**	**RF**	**LASSO**	**RF**	**RNN**
H1	0.7	0.5	0.1	0.5	0.3	0.5	0.67
H2	0.2		0.5		0.6		
H3	0.5		0.3		0.4		
L1	0.3		0.7		0.7		
L2	0.6		0.1		0.3		
L3	0.2		0.1		0.3		

we inspected the per-position breakdown of the actual and predicted paratopes (Figures 5.13 and 5.14), the behaviours of these models were similar to the patterns seen on the original Parapred set. In CDRH1, considering the neighbourhood contribution (*i.e.* the "Triplet" featurisation) tended to under-predict while the LASSO Single Meiler model over-predicted on the latter half of the loop along and coincided with Parapred. In CDRH2, all of our models matched the behaviour of Parapred. In CDRH3, Parapred outperformed all of our models, except on positions 111E and 111F that Parapred failed to predict. The overall behaviour of the LASSO

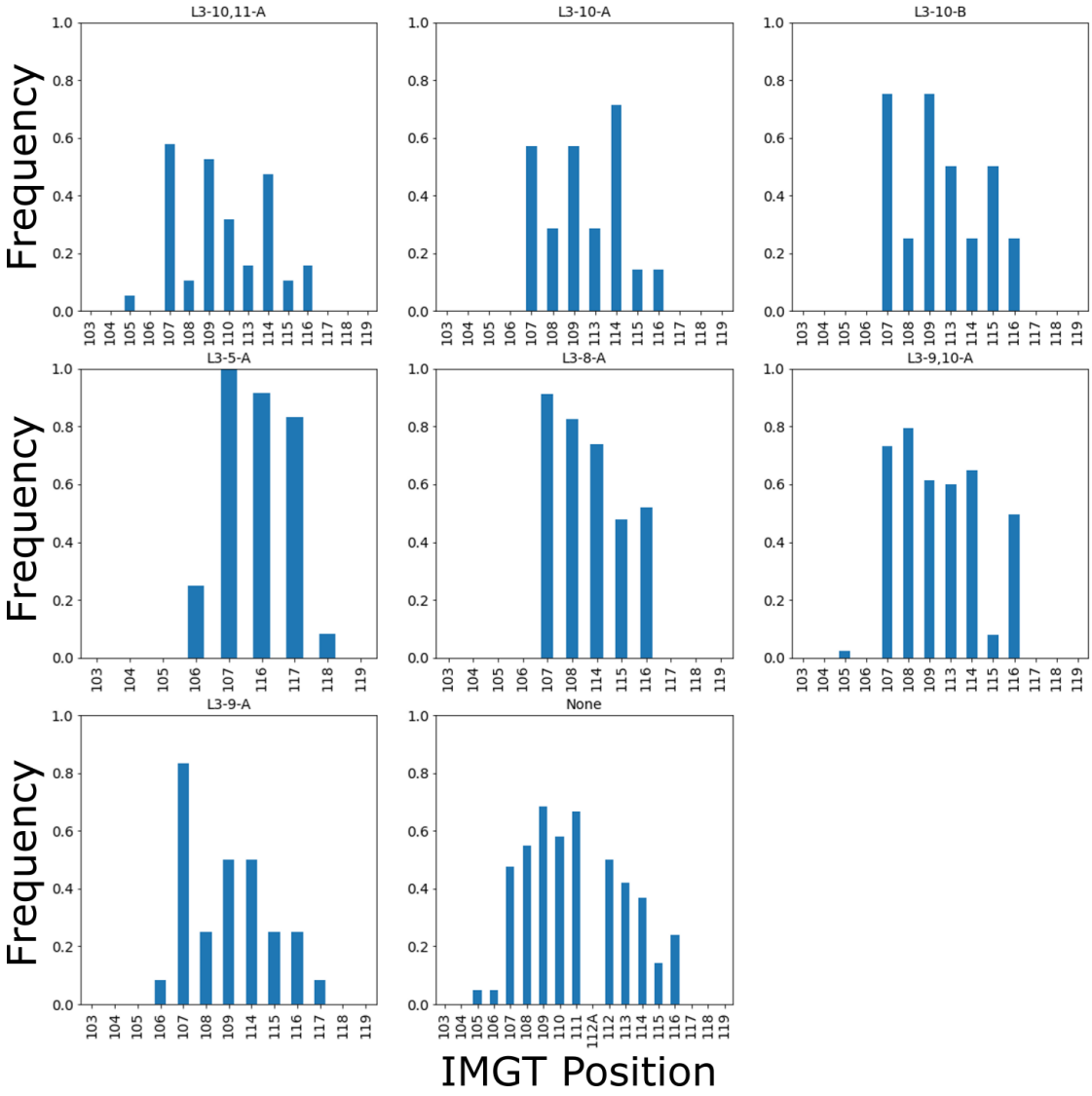


Figure 5.8: Relative frequencies of the actual paratopes by IMGT-positions, stratified by CDRL3 canonical forms in the Parapred set. See Figure 5.3 for details.

Triplet Meiler appeared to best mimic that of Parapred. Similarly in the light chain CDRs, our RF Single or Triplet Meiler models were mostly closely matched to the Parapred predictions and actual paratope propensity in CDRL1. In CDRL2, our models tended to under-perform, and so did Parapred. CDRL3 saw a slight decrease in performance in our models. LASSO Single and Triplet Meiler models jointly approximated Parapred's predictions.

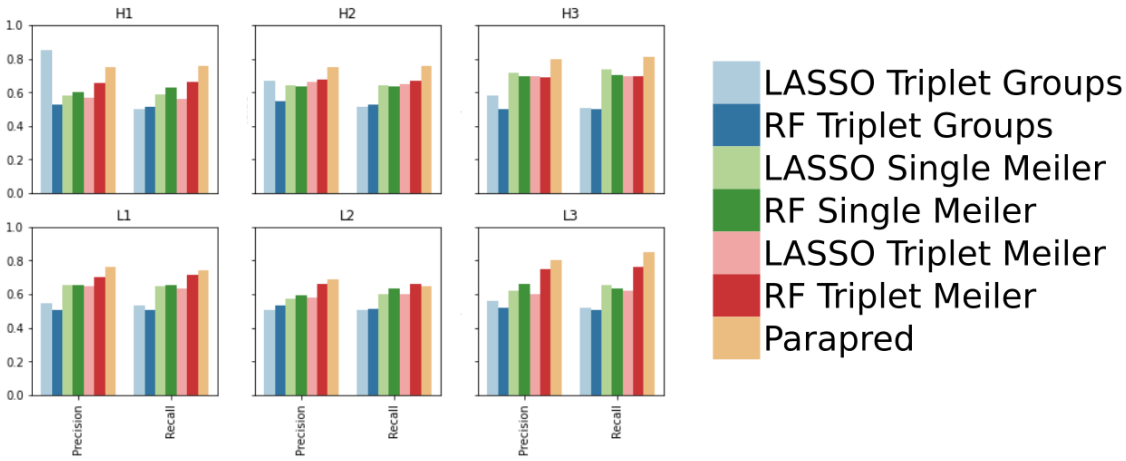


Figure 5.9: Performance of the models by CDR types, for the Parapred set. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

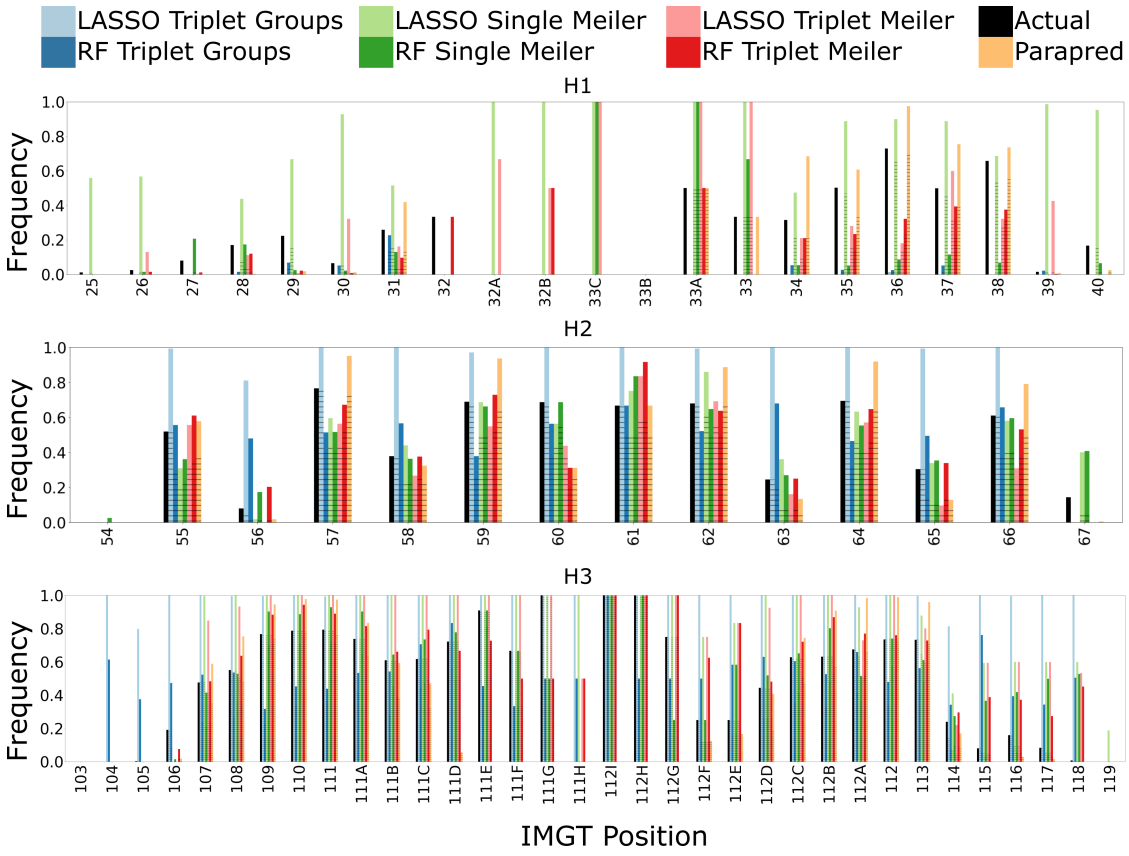


Figure 5.10: Relative frequencies of the actual and predicted paratopes on heavy chain CDRs in the Parapred set, by IMGT-positions. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

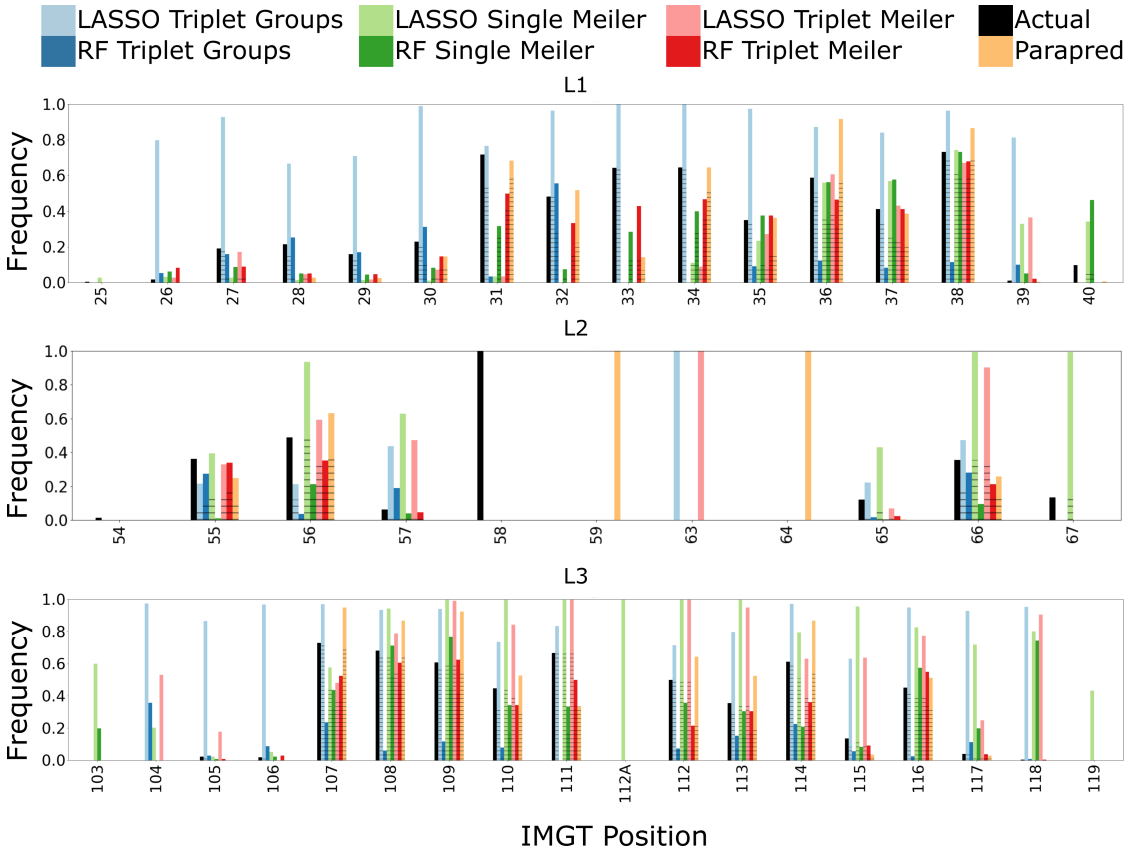


Figure 5.11: Relative frequencies of the actual and predicted paratopes on light chain CDRs in the Parapred set, by IMGT-positions. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

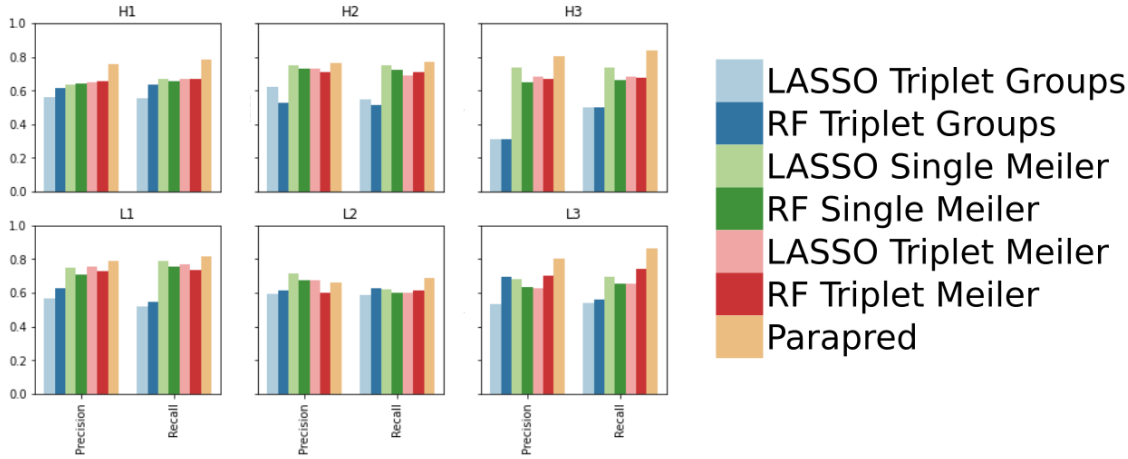


Figure 5.12: Performance of the models by CDR types in the blind set. See Section 5.2.3.1 and Table 5.2 for the nomenclature of the models in the legend.

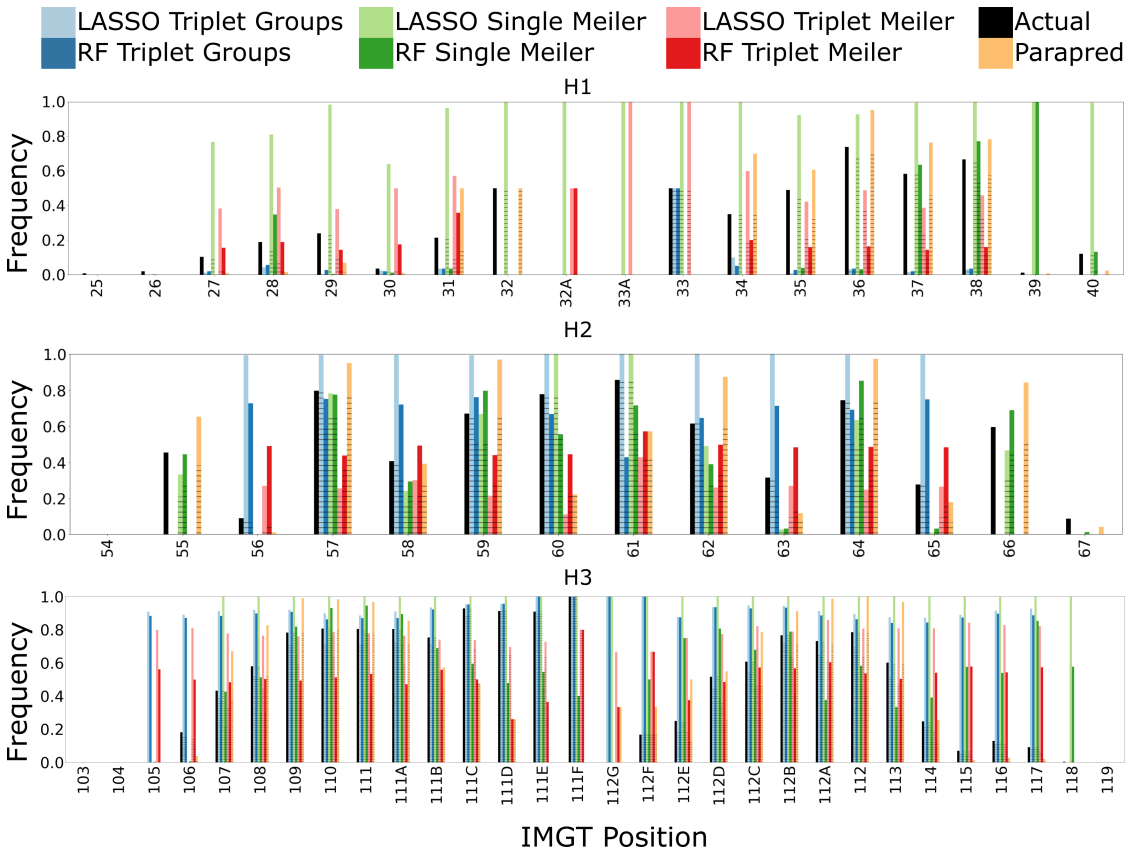


Figure 5.13: Relative frequencies of the actual and predicted paratopes on heavy chain CDRs in the blind set. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

5.3.4 Stratifying the training set by sequence lengths

Since the CDR sequences in structural datasets are shorter than those found in repertoire data, we simulated a situation where only the shorter loops were used to train the models and tested if these models were able to perform well on the longer loop lengths. This is a very small test set as few IMGT-defined CDR loops were longer than 14 residues, except in CDRH3. Figure 5.15 shows that our models have similar precision and recall as Parapred. On the positional frequencies of the actual and predicted paratopes (Figure 5.16), the general behaviour is similar to the cross-validation and blind set tests reported above: the middle sections (positions 111B – 112B) of these longer CDRH3 loops were not predicted by Parapred trained on short loops only. Our RF Triplet Meiler model exhibited similar behaviours in the length-stratified set as in the original Parapred set, while our other models

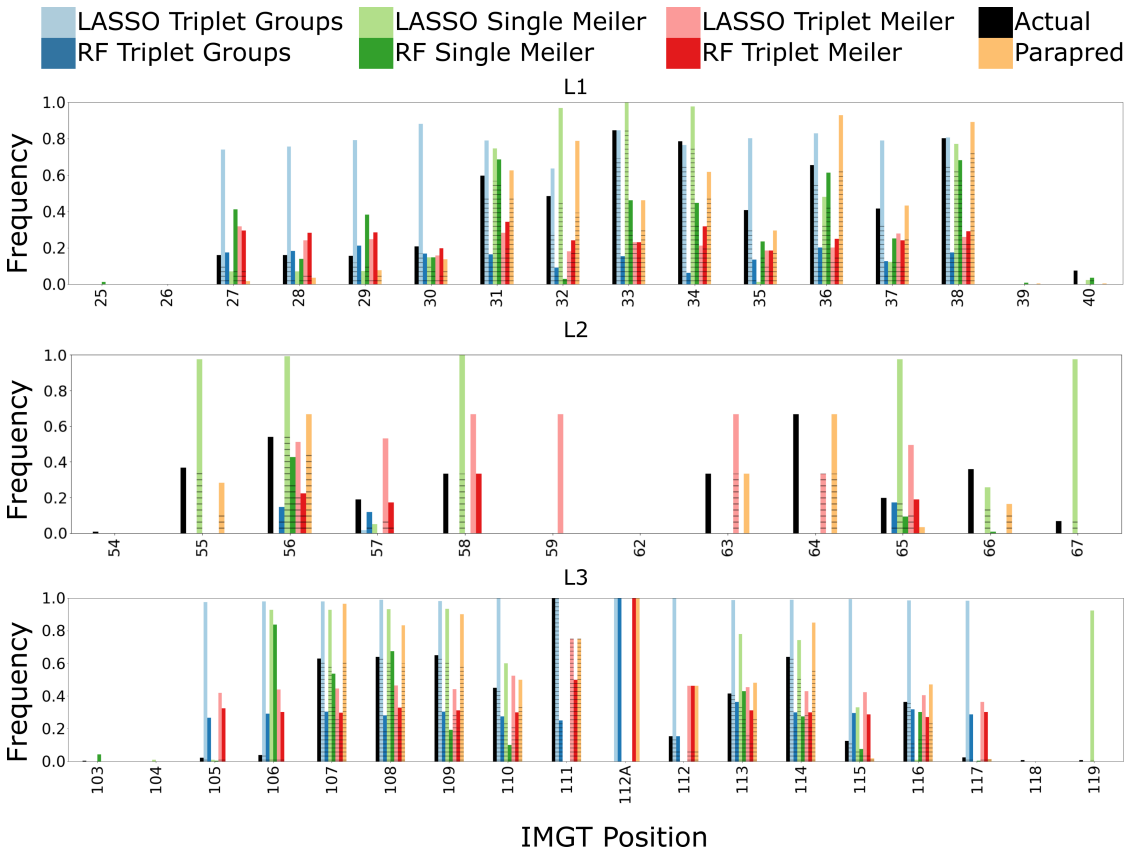


Figure 5.14: Relative frequencies of the actual and predicted paratopes on light chain CDRs in the blind set. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

saw a slight drop in performance. Even so, our RF models appeared robust when predicting on the unseen sequence lengths: they were able to annotate the centre regions of long loops correctly in most cases. This implies our models had lower dependence on the sequence lengths and were able to infer paratope patterns in long loops even when no long loops had been used for training.

5.3.5 Peptide-binding paratopes

We inspected the applicability of the modelling frameworks on paratopes against a different antigen type – peptides. Using the identical frameworks, we trained the seven models (Parapred and our six featurisation-model combinations) on peptide-binding antibodies. The CDR-specific thresholds for the LASSO models

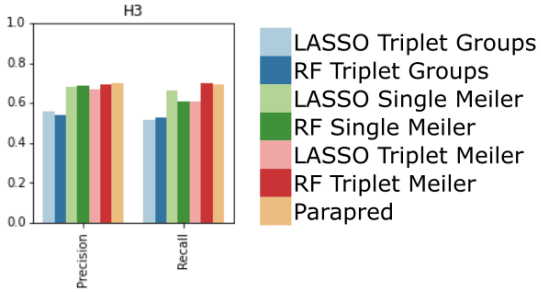


Figure 5.15: Performance of the models by CDR types, in the length-stratified set. See Section 5.2.3.1 and Table 5.2 for the nomenclature of the models in the legend.

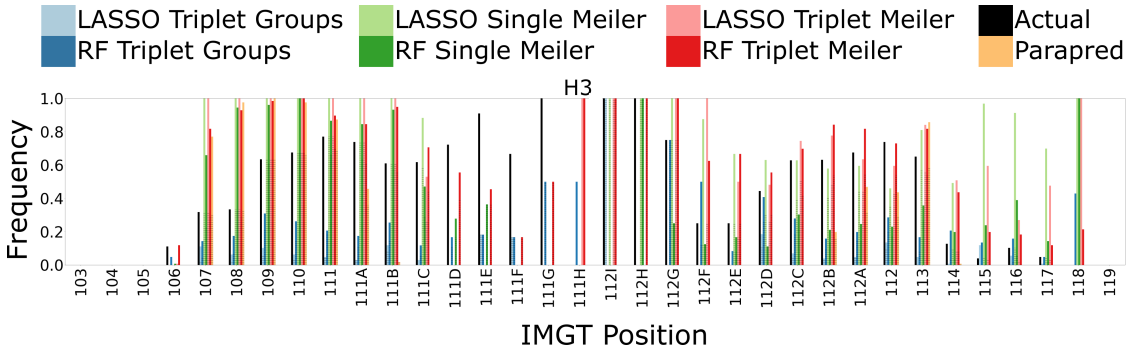


Figure 5.16: Relative frequencies of the actual and predicted paratopes on the CDRs, by IMGT-positions, in the length-stratified set. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

were selected with the same procedure as before by maximising the MCC (see Table 5.3). Some of the selected thresholds were relatively high (0.8 – 1.0). At these thresholds, very few predictions were made, particularly in CDRL3 across all featurisation schemes and the Triplet Groups tokens for all CDR types. This was a result of the threshold selection step not being able to find a balance between the false positive and false negative rates. This suggests that this featurisation method with a LASSO model might be unsuitable for some of these tasks.

Figure 5.17 shows similar results to protein-binding antibodies (Figure 5.9), that Parapred performed consistently better than our featurisation-model combinations. On the positional frequencies of the actual and predicted paratopes, the lack of predictions around the N-terminus of the CDRH1 and CDRL1 shown in protein-binding antibodies was recapitulated in the peptide-binding set (Figures 5.18 and

5.19). Observations were consistent between protein-binding and peptide-binding CDRH2, L2 and L3. On CDRH3, the behaviour on the peptide-binding paratope predictions was slightly different for Parapred, where the middle section was not totally missed out (Figure 5.18). It appeared that peptide-binding antibodies had slightly shorter CDRH3 than protein-binding antibodies, thus the Parapred predictions were more robust.

Table 5.3: Selected thresholds for the peptide set. Refer to Table 5.2 for the model nomenclature.

Featurisation	**Triplet Groups**		**Single Meiler**		**Triplet Meiler**		**Parapred**
Model	**LASSO**	**RF**	**LASSO**	**RF**	**LASSO**	**RF**	**RNN**
H1	1.0	0.5	0.3	0.5	0.1	0.5	0.67
H2	0.5		0.5		0.5		
H3	0.8		0.4		0.4		
L1	1.0		0.8		0.2		
L2	1.0		0.2		0.2		
L3	0.8		1.0		1.0		

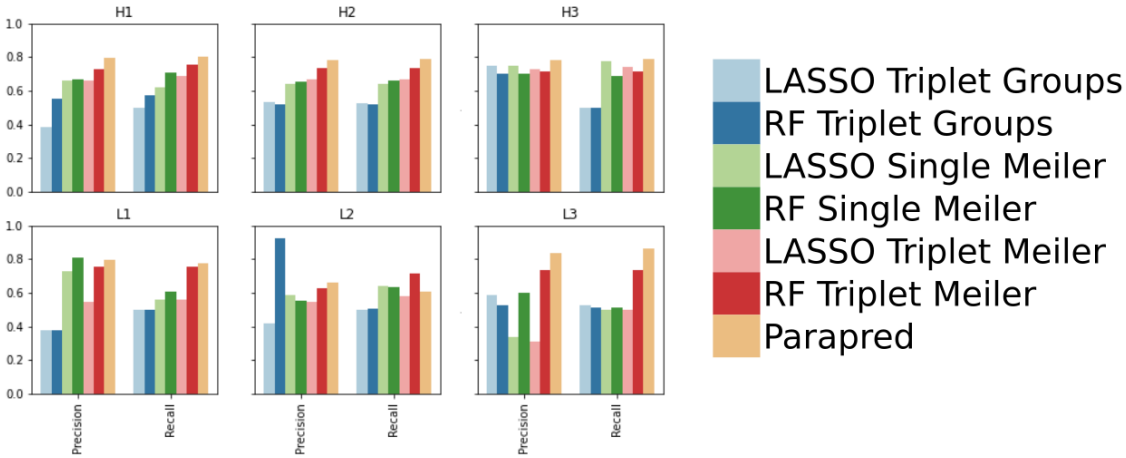


Figure 5.17: Performance of the models by CDR types in the peptide set. See Section 5.2.3.1 and Table 5.2 for the nomenclature of the models in the legend.

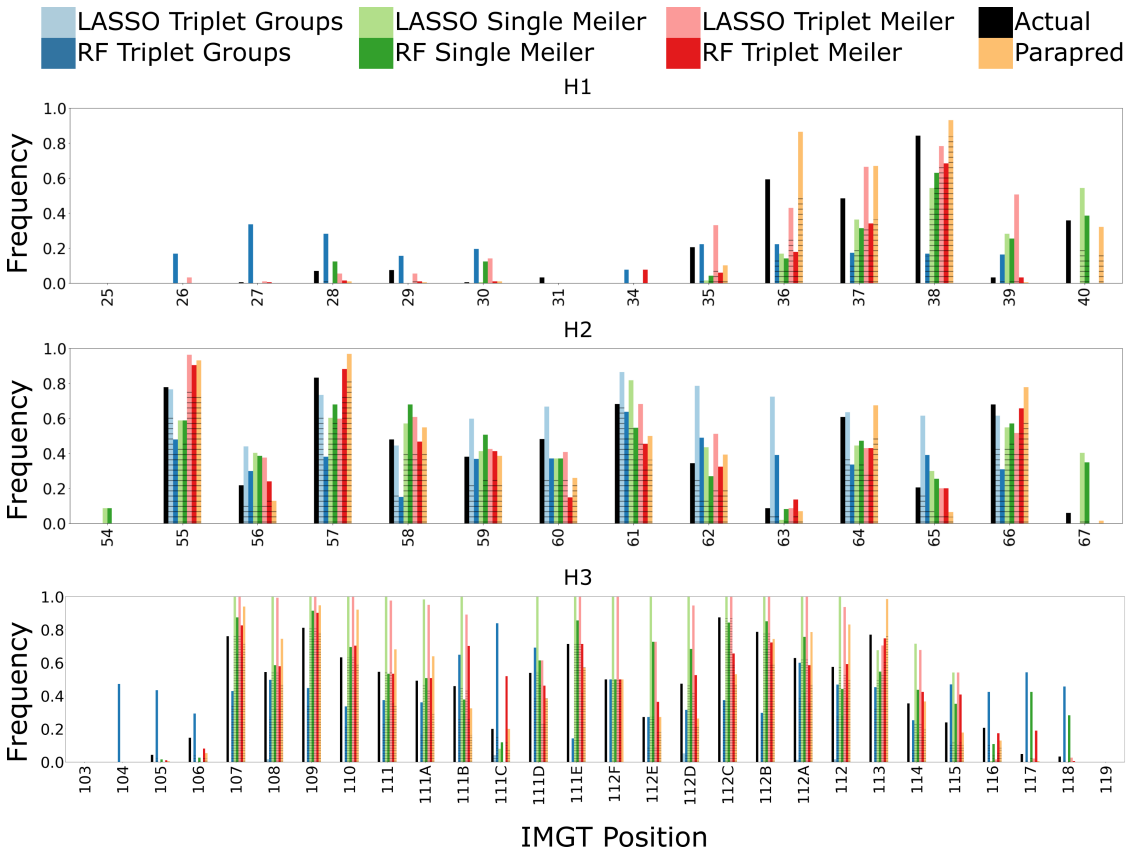


Figure 5.18: Relative frequencies of the actual and predicted paratopes on heavy chain CDRs in the peptide set, by IMGT-positions. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

5.3.6 Feature importance in protein-binding and peptide-binding antibodies

Since in most cases, the RF Triplet Meiler model was closest to Parapred's predictive ability, it may be used (because of its interpretability) to examine the features used in paratope prediction. We compared the feature ranking from this model, between the protein-binding and peptide-binding antibodies. Figure 5.20 and Appendix Figures D.1 – D.5 show the feature importance ranking by CDRs in all six models, coloured by the feature types. In most of the CDRs, the top-ranked features in the RF Triplet Meiler model were related to their positions (Figure 5.20). This was particularly obvious in CDRH3 where the top 20 features were mostly the positions (in green). In canonical CDRs, only some of the canonical forms (in orange) were predictive. For instance, in CDRH2, three of the five canonical forms were among

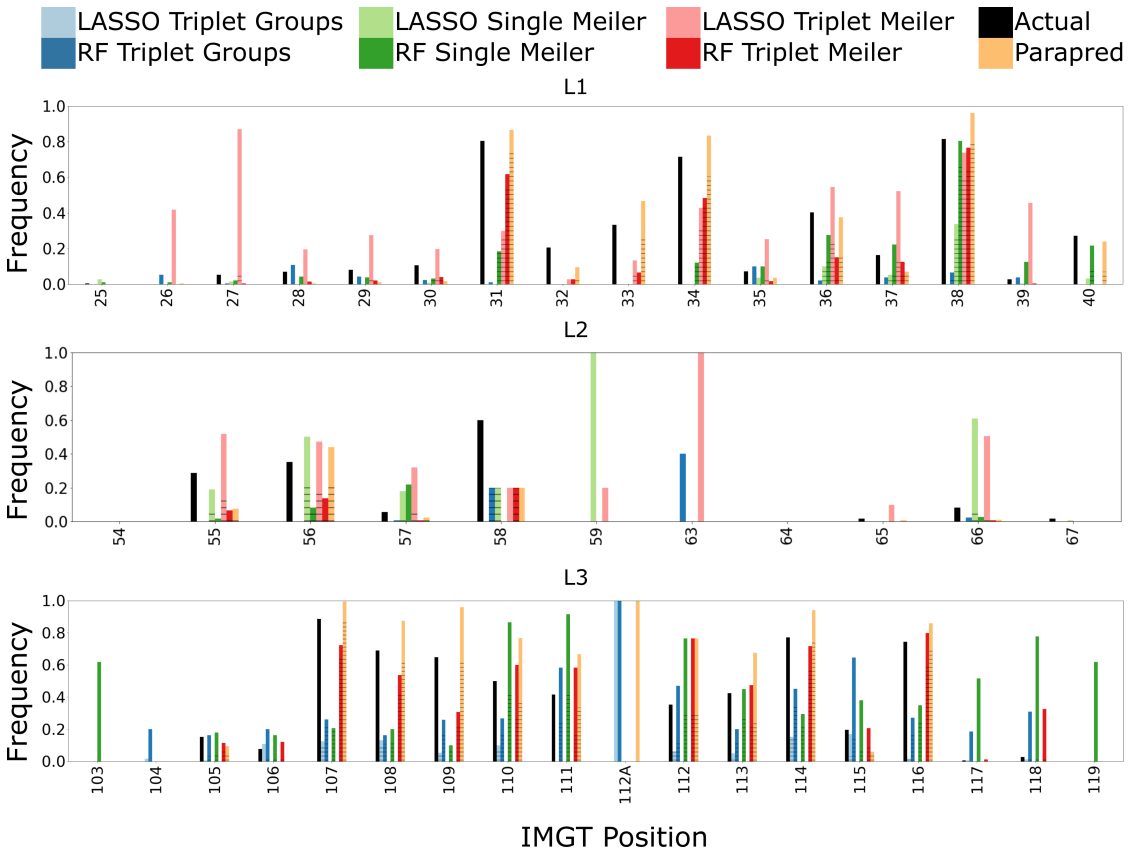


Figure 5.19: Relative frequencies of the actual and predicted paratopes on light chain CDRs in the peptide set, by IMGT-positions. The portion of true positives is lightly shaded. Refer to Section 5.2.3.1 and Table 5.2 for model nomenclature.

the top 10 important features in the RF Triplet Meiler model, but in the other CDR types, fewer canonical forms had a high ranking.

Comparing between continuous or categorical physicochemical features, the feature ranking plots of the Triplet Groups featurisation (Appendix Figures D.1 and D.2) show that the models had placed heavy weights on the position and less on the physicochemical features. This suggests, the large number of one-hot-encoded physicochemical features might have created a sparse matrix as the input, that could be difficult for the statistical models without any dimensionality reduction. This has potentially skewed the feature importance in favour for the denser regions represented by the residue positions.

In the neighbourhood consideration, Single Meiler models showed similar rankings to their Triplet counterparts (Figure 5.20 and Appendix Figures D.3 – D.5). The additional residue information from the sequential neighbours occasionally added more noise to the linear model or decision three, though the overall trend of the additional features roughly followed that of the main trend.

Overall, protein-binding and peptide-binding paratopes appeared to use different patterns for binding. The Spearman's rank correlation ρ of these plots are sometimes above 0.70. In Figure 5.20, CDRL2 has a high ρ but may not be as significant due to its short length, higher sequence conservation across antibodies, and a consistently low binding likelihood. CDRH1 and L3 have ρ above 0.70, indicating that both of these loops may use similar patterns to engage protein and peptide antigens. However in the other CDR types, the low values of ρ suggest that the correlations between the feature importance in both types of binders were weak, thus their paratopes had different features.

5.4 Discussion

In this chapter, we evaluated Parapred and observed its lack of prediction in the centre of long loops. We identified three categories of features that could be used with simple statistical models to achieve similar performance while filling the gaps in Parapred predictions. To simulate a real-life scenario where the predictors are applied on sequencing datasets with longer and more diverse CDRs than that of its structural training set, we stratified the original training set by length. We found that our simpler models were better at generalising across sequence lengths than Parapred. We assessed the applicability of these models on peptide-binding antibodies, and they performed similarly to the protein-binding set. We compared the differences in the key predictive features in protein-binding and peptide-binding antibodies. The ranking indicated that paratopes against these two types of antigens had different features, but residue positions were ranked among the most important

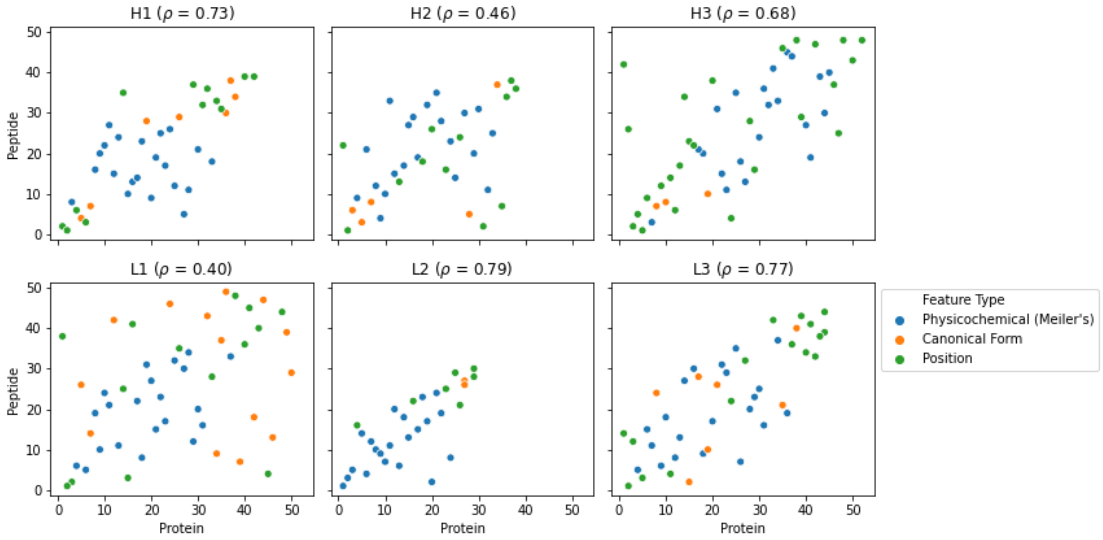


Figure 5.20: Feature importance of the the RF Triplet Meiler model by CDR types, between protein-binding (Parapred set) and peptide-binding (peptide set) antibodies. Each point represents a feature, coloured by whether it is a physicochemical feature (blue), canonical form or sequence length in CDRH3 (orange) and the position (green). The x- and y-axes are the feature ranks in the protein and peptide sets. The bottom-left region houses the top-ranked features in both sets. ρ is the Spearman's rank correlation coefficient between the features from the models trained on protein-binding and peptide-binding paratopes.

features in both sets.

This work only explored one way of categorising residues and seven continuous values characterising residue chemical properties. Alternative featurisation strategies can be used to enrich the feature space, including one-hot-encoding by residues, different tokenisation schemes (Pommié et al., 2004) and chemical features (Kawashima et al., 2007). Furthermore, the structural information is only limited to the canonical CDRs. To expand on the structural features, an antibody model is needed along with a descriptor that can turn the paratope structures into a feature column as input to the model. Hashes in Antibody i-Patch (Krawczyk et al., 2013) and Ab-Ligity (Chapter 4), and other surface descriptors (mentioned in Section 1.4.2.2) may be incorporated to render the spatial correlation between residues from different CDRs. Additional tuning may be necessary to compensate

for the errors in structural modelling.

Exploring the neighbourhood contribution is crucial as paratope residues work collectively. The feature importance recovery in our tools provides the scope to study feature correlations in the triplets, CDR sequences and with the neighbouring CDR loops. The Triplet encoding would also benefit from using multi-output columns for each of the residues to more holistically infer and assess the contribution of sequential neighbours in binding. Both of these tests would draw insights to supplement the current repertoire analysis that primarily focusses on the dynamics in germline usage and sequence convergence (Galson et al., 2015), but lacks an emphasis on the actual binding site selection and affinity maturation.

The transparency of our linear and ensemble models enables us to compare the feature importance in protein- and peptide-binding antibodies. We found that in both cases, the residue positions play an critical role in the prediction of the models. This was also observed by Olimpieri et al. (2013) where the most important variables for paratope predictions included the germline, antigen volume and predominantly residue positions. The feature ranking differences found in our work between paratopes against the two antigen types could be attributed to the binding modes. Peptides and smaller antigens can bind deeper in a "pocket", while protein antigens typically bind to a flatter surface on the paratope (Lee et al., 2006). In these scenarios, residues in the centre of and those closer to the two termini of the CDR might adopt different binding preferences when engaging protein and peptide antigens. Further analysis of the existing pipeline may potentially validate these observations.

Due to the different inclinations for CDR sequence lengths in the repertoire and structural datasets, a major focus of this work was to assess if the prediction models trained on the structural datasets were overfitted to the shorter lengths. We therefore simulated the situation by stratifying the training set by sequence lengths

and observed any length dependence. Our featurisation methods showed less of such dependence than Parapred. This could be an advantage when extrapolating beyond the structural datasets, in a scenario where we leverage repertoire data to study binding site characteristics and diversity. For instance, Richardson et al. (2020) used Parapred to annotate the paratopes in sequencing datasets and clustered the antibodies by their predicted paratope identities. The missing paratope predictions in the middle of long loops could adversely impact the ability to draw conclusion on the binding site diversity in repertoires.

5.5 Chapter Summary

We conducted an appraisal of the sequence-based deep learning paratope predictor, Parapred, and proposed potential predictive paratope features including the residue position, physicochemical properties, neighbourhood and canonical forms. We built linear and ensemble models to capture these features, but the models proposed did not outperform Parapred in overall precision and recall. We verified that Parapred had limited performance on long CDR sequences, that was in part alleviated by our simpler models. We further retrained the models on the peptide-binding paratopes and compared the feature importance ranking in peptide-binding against protein-binding paratopes. We found that paratope patterns were different when they engaged antigens of different types, but the residue positions remained the top-ranked features.

In the next chapter, we will summarise the findings in these four result chapters and the thesis. We will also outline potential extensions to the analyses and tools proposed in these four chapters. The challenges and perspectives on binding site analysis across repertoires will conclude the thesis.

6 Future work and Conclusions

Contents

This chapter summarises the four result chapters and suggests potential extensions from the work. It also highlights recent developments and the future directions related to this work, and the key steps to bridge the gap between the volume and depth of repertoire sequencing and the fine details in structural data.

6.1 Chapter Conclusions

The aim of the thesis was to study the binding site structures of antigen receptors. From the structural analysis, we built tools that can be applied to sequencing datasets to leverage their volume to improve our understanding of binding site variability in the natural repertoire.

6.1.1 Chapter 2

In this chapter, we built upon an existing structural clustering method developed by Nowak et al. (2016), that identifies CDR canonical forms. We then built a canonical form predictor (SCALOP) that used position-specific scoring matrix (PSSM) to capture the sequence pattern in each canonical form. We compared its performance to an existing database-search loop modelling tool FREAD (Deane and Blundell, 2001; Choi and Deane, 2010). The precision and recall between the two methods were similar on a cross-validation set. On a large sequencing dataset, we found that SCALOP and FREAD made consistent predictions, but the coverage of SCALOP was slightly lower than that of FREAD. The major advantage of SCALOP is its fast run-time – it is 800-times faster than FREAD, enabling it to rapidly annotate large sequencing datasets. Since the development of SCALOP for antibody CDRs, it has been integrated into the SAAB+ pipeline (Kovaltsuk et al., 2020) to speed up the annotation of the canonical CDRs in large datasets of sequences.

As the structural data expands over time, we also investigated the changes in back-dating the training set. We found that the PSSM algorithm was robust to the constant renewal of the underlying data. In addition, we observed an increasing number of canonical forms as more data had been added over the years, and concluded the need to automatically update the SCALOP database. Finally we applied SCALOP on a mutagenesis dataset (Adams et al., 2016) and demonstrated that SCALOP was able to quickly highlight mutants that switched canonical forms and due to this had weaker affinity for their target.

6.1.2 Chapter 3

As antibodies and TCRs share similar genetic mechanisms and constructs, we compared their binding sites to try and identify features that lead to their biological functions. We ran the SCALOP pipeline on TCR CDRs and established their

structural clusters and the associated PSSMs. The analogous canonical CDR types from the two antigen receptors (*i.e.* CDRH1/β1, CDRH2/β2, CDRL1/α1, CDRL2/α2 and CDRL3/α3) were contrasted. Our analysis showed that antibodies and TCRs tend to have different CDR loop lengths. Where they share similar sequence lengths, they adopt different conformations. In the two cases (one in L1/α1 and one in L3/α3) where loops from antibodies and TCRs were clustered together, they used different sequence patterns to encode for the same backbone shape. For the non-canonical CDRH3/β3 comparison, we found that the majority of antibody CDRH3 loops had a kinked torso, while over 97% of TCR CDRβ3 loops had an extended base. Overall, these results suggest that antibody and TCR CDRs are distinct from one another.

We further inspected the structural variability in antibody and TCR CDR loops. Of the unique CDR sequences with multiple example structures in the SAbDab or STCRDab set, only 8% of antibody CDR loops adopt multiple distinct conformations, in contrast to $\sim$20% in TCR CDRs. Consistent with earlier molecular dynamic studies on specific case studies, these results suggest greater binding site flexibility in TCR which may be correlated to their polyspecificity.

Given the higher structural variability in TCR binding sites was not captured by any existing single-state TCR modelling tools, we adapted the antibody modelling pipeline, ABodyBuilder (Leem et al., 2016), for TCR structural modelling. The performance of this new tool, TCRBuilder is comparable to existing tools, with the additional benefit of returning a multi-state TCR model that collectively represents the TCR binding site conformations that allowed it to engage multiple different conformations.

6.1.3 Chapter 4

Sequence similarity is typically used as a clustering metric to identify antibodies with similar binding modes. However, this method neglects sequence-dissimilar but structurally similar binders that engage the same epitope. We developed Ab-Ligity, a structure-based metric that is able to capture and compare the spatial and physicochemical information of the antibody-antigen binding sites. Ab-Ligity is designed to work on antibody model and predicted paratopes. We proved that Ab-Ligity rapidly and accurately identifies sequence-dissimilar paratopes against the same epitope.

As the majority of the available repertoire sequences are unpaired heavy or light chains, we examined if Ab-Ligity and InterComp (an existing tool for general protein surface comparison) were able to work on only the heavy or light chain of antibodies. We found that performance was impacted for both methods. Ab-Ligity saw a slight drop in the precision and recall in both VH-only and VL-only paratope comparisons, while InterComp suffered more adversely from the small surfaces in the VH-only or VL-only paratope surface patches.

6.1.4 Chapter 5

The final result chapter showed the preliminary results of our appraisal of the neural network-based paratope predictor, Parapred (Liberis et al., 2018), and proposed alternative, interpretable featurisation schemes and statistical models for paratope prediction. We first showed the failure of Parapred to recognise the centre of long loops as paratopes. Then we used the position, canonical form or sequence length, physicochemical features and neighbourhood to featurise CDR sequences. We used two simple models, LASSO and random forest. On the cross-validation and blind sets, these feature-model combinations achieved similar but slightly lower precision and recall compared to Parapred. However, our models were able to make predictions in the centre of long loops in the full and length-stratified

sets, suggesting that our models were potentially less reliant on the sequence lengths.

6.2 Future Work

The advent of new experimental techniques and novel antibody formats has created opportunities for building alternative computational models. Below we outline the future work that could be done to further utilise the tools presented in this thesis, and possible orthogonal ways of studying paratope-epitope interactions.

6.2.1 TCR binding sites

We showed that binding sites in TCR are structurally more variable than those in antibodies despite their similar genetic mechanisms and folds, and subsequently built a multi-state TCR modelling tool. To further understand the mechanism of polyspecificity in TCRs, it is crucial to model the TCR-pMHC complex. Due to the largely invariant pMHC conformation, template-based complex modelling has typically been used (*e.g.* Pierce and Weng, 2013; Hoffmann et al., 2018; Jensen et al., 2019). However, they provide only a single snapshot and do not capture the polyspecific binding in TCRs. TCRBuilder provides a basis for capturing the alternative binding configurations of a TCR, thereby allowing multiple starting TCR structures for docking, template modelling or further energetic refinement.

6.2.2 Epitope-paratope searching pipeline

The hash tables generated by Ab-Ligity to describe paratopes and epitopes in the set of known antibody structures can be used to create a "dictionary" of paratope-epitope pairs. Given a query epitope, such a dictionary can in theory suggest the complementary paratope features, that can in turn be used to identify possible alternative binders within a sequencing dataset. The latter can be from a single sequencing study, or from a pre-curated set of antibody structures. Based on a set

of public antibody structures commonly found across multiple patients' immune repertoires (Raybould et al., 2020), it may be possible to pre-build a database of Ab-Ligity hash tables describing the paratopes represented in these "public" structures. Combining both the paratope-epitope dictionary and a pool of public structures, the pipeline can possibly support virtual screening by running a search through the paratope-epitope dictionary, and conduct an *in silico* hit expansion within the curated screening library of antibodies.

The application of this approach is bounded by the small coverage in the structural space of antibody-antigen interfaces in the PDB. A possible way to generate more data for this dictionary is to leverage partner-specific binding interface predictions. A number of antibody-specific epitope prediction tools (*e.g.* Sela-Culang et al., 2015) have been developed using graphical models and neural networks. In addition, leveraging complex modelling or macromolecular docking, we may be able to estimate how an antibody would be perceived to engage with a potential epitope. These algorithms may potentially generate additional artificial data to expand the dictionary. Admittedly all of these tools are predictions and have their inherent errors and discrepancies from the true binding mode, and will have to be thoroughly evaluated for the signal-to-noise trade-off they bring to expand the dictionary.

6.2.3 Other protein-protein interfaces and consensus analysis

Moving beyond antibody-antigen interfaces comparison, a potential extrapolation would involve testing the algorithm on general protein-protein interfaces, as illustrated in InterComp (Mirabello and Wallner, 2018). Likewise, postulating the paratopes of non-protein binding antibodies may require an intermediate version of the original pharmacophore-based Ligity and the current residue-based approach.

One of the earlier surface clustering methods applied to antibody-antigen interfaces, pCLICK (Nguyen et al., 2011, 2017), demonstrated that overlaying similar paratope surfaces might highlight the important commonality and infer the key binding residues. By expanding the protein-protein interface database, it may be possible to identify consensus interaction patterns in homologous proteins with their native binding partners. In turn, this may inform the key interactions that shall exist in the antibody to mimic the native protein-protein interactions, and those that need to be displaced to destabilise the native binding.

6.2.4 Leveraging epitope mapping and mutagenesis datasets

Synthetic biology and laboratory automation have given rise to deep mutagenesis capabilities in various research groups (*e.g.* Koenig et al., 2015; Adams et al., 2016; Koenig et al., 2017; Warszawski et al., 2019; Lim et al., 2019; Olson et al., 2019). Most of these reported their binding affinity upon mutations (Koenig et al., 2015; Adams et al., 2016; Koenig et al., 2017; Warszawski et al., 2019; Lim et al., 2019), but some also assessed thermostability and other fitness parameters (Warszawski et al., 2019; Olson et al., 2019). These analyses have inspired computational tools that use sequence-based, simple statistical models such as PSSM to evaluate the interfacial energetic changes introduced by the mutations, thus informing interface design (*e.g.* Warszawski et al., 2019). These studies were typically constrained to a single target. To understand the general principle behind interface complementarity, the ideal scenario is to aggregate data from multiple studies. Along a similar line of thought, AB-Bind (Sirin et al., 2016) was set up to store the affinity measurements of the existing set of antibody structures and some of their mutants. However, most of the deep sequencing datasets do not have a solved crystal structure for the wild type antibody and were therefore excluded from this database. An extension to current binding affinity databases including sequence-affinity mapping would provide a comprehensive view of what the different experimental techniques have

unveiled for each of the targets and antibody libraries.

This type of data storage has been used extensively in small molecule discovery and has proven its usefulness in understanding protein modulation and aiding compound design (Gaulton et al., 2017). We envisaged the curation of an analogous database would enrich the current modelling approach by integrating multiple data sources against different targets.

6.2.5 Integrative studies on specific targets

The COVID-19 pandemic has seen considerable effort being expended into understanding the biology of the disease and searching for potential cures. In particular, the rapid speed-up in clinical-experimental collaborations has released sets of patients' post-infection repertoires, complemented by epitope mapping assays and crystal structures. This could be among the most extensively-studied systems, exceeding even HIV and flu, with the advantage of good epitope coverage on the main target protein. The challenge is to assimilate multiple data sources into a sensible pipeline to identify the parts of the repertoire representing pre-existing immunity, to reduce complexity by sequence clustering without scrambling the weak signal that defines specificity, and to decipher the evolutionary code that renders immunity in some patients but not all. Germline, sequence features and structural complementarity play an important role in this analysis and have been partly covered by the multiple types of experiments with published data. This extensive dataset may enable us to study antibody-antigen binding and understand acute immune response.

6.3 Closing remarks

In this thesis, we presented our analysis on antigen receptor binding sites. We compared the structural differences in the CDRs of the two antigen receptors. TCR

CDRs were found to be more structurally variable than their antibody counterparts. To capture these observations, we developed SCALOP for antibody and TCR CDRs respectively, and TCRBuilder – a multistate TCR modelling tool. Furthermore, to identify sequence-dissimilar antibodies that engage the same epitope, we built Ab-Ligity, a structure-based metric that quantifies the similarity between binding sites. Finally, we studied the paratope features and proposed a number of simpler featurisation schemes and statistical models to predict paratopes. We observed that protein-binding and peptide-binding paratopes have different features but residue position tends to be the most predictive features for paratopes.

Appendices

A

SCALOP Appendix

A.1 Summary of clusters

A.1.1 Overview of clusters in each CDR type

Here we present a summary of the structural clusters found in each CDR type.

Table A.1: Summary statistics of clusters in each CDR type.

CDR	Total number of sequences	Clustering Threshold (Å)	Portion of clustered sequences	Number of clusters
H1	2747	0.80	81.03%	3
H2	2819	0.63	83.29%	4
L1	2605	0.82	92.32%	12
L2	2765	1	98.41%	1
L3	2713	0.91	83.27%	7

Table A.2: Cluster-specific details of each CDR type.

CDR	Clusters	Lengths	#Unique	#Redundant
H1	H1-13-A	13	605	2047
	H1-14-A	14	20	80
	H1-15-A	15	20	99
	#Clustered	**2226**	**#Total**	**2747**
H2	H2-9-A	9	170	608
	H2-10-A	10	366	1001
	H2-10-B	10	187	561
	H2-12-A	12	39	178
	#Clustered	**2348**	**#Total**	**2819**
L1	L1-10-A	10	26	80
	L1-11-A	11	243	1051
	L1-11-B	11	40	120
	L1-12-A	12	24	54
	L1-12-B	12	11	87
	L1-13-A	13	33	95
	L1-13-B	13	9	47
	L1-13-C	13	7	20
	L1-14-A	14	21	91
	L1-14-B	14	11	100
	L1-15-A	15	40	97
	L1-16,17-A	16,17	144	563
	#Clustered	**2405**	**#Total**	**2605**
L2	L2-8-A	8	449	2721
	#Clustered	**2721**	**#Total**	**2765**
L3	L3-5-A	5	12	49
	L3-8-A	8	40	141
	L3-9-A	9	29	141
	L3-9,10-A	9,10	470	1729
	L3-10-A	10	20	92
	L3-10-B	10	8	10
	L3-10,11-A	10,11	41	97
	#Clustered	**2259**	**#Total**	**2713**

A.2 Cross-validation threshold

Table A.3: F1 score from the cross-validation. The highlighted cells indicate the maximum F1-score across the different thresholds for each CDR.

Assignment Thresholds	H1	H2	L1	L2	L3
-2	0.9337	0.9661	0.9765	0.9956	0.9541
-1.5	0.934	0.9661	0.9765	0.9956	0.9601
-1	0.9338	0.966	0.9765	0.9956	0.9619
-0.5	0.9342	0.9664	0.9771	0.9945	0.9615
0	0.9363	0.9639	0.9755	0.9923	0.961
0.5	0.9408	0.9526	0.9736	0.9811	0.9571
1	0.9333	0.9259	0.9663	0.9521	0.9347
1.5	0.9135	0.8324	0.9537	0.8649	0.8818

B
TCR Appendix

B.1 TCR canonical forms

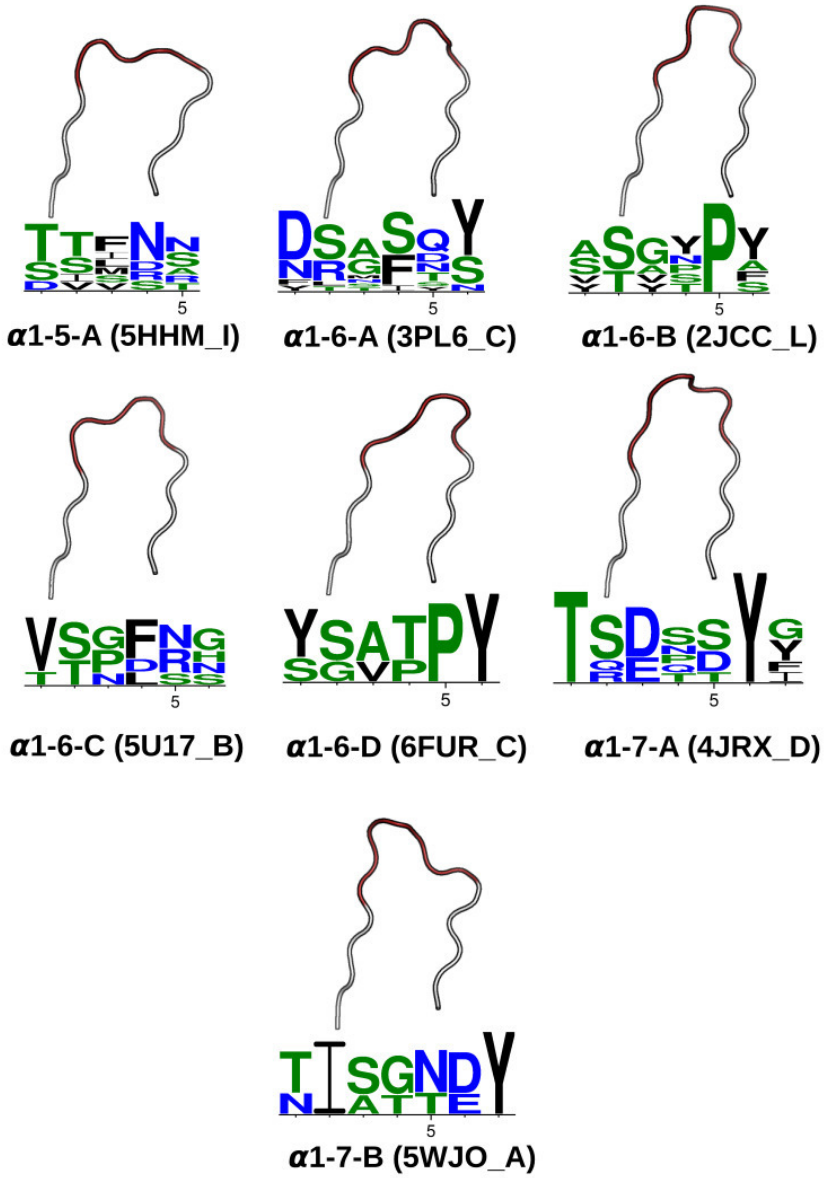


Figure B.1: CDRα1 loop clusters. At a 1.0Å clustering threshold, our DBSCAN method identified seven canonical classes (see Section 3.2.4). Each class has at least five structures and two unique sequences. Anchors are coloured white, while the CDRα1 region (IMGT 27–38) is coloured. The Protein Data Bank (PDB; Berman et al. 2000) four-letter code and the chain identifier of the centroid structure is shown in the bracket next to the cluster name. The sequence pattern below each centroid structure is generated by WebLogo (Crooks et al., 2004), using the unique sequences of the cluster: α1-5-A has 7, α1-6-A has 13, α1-6-B has 6, α1-6-C has 5, α1-6-D has 3, α1-7-A has 6 and α1-7-B has 3.

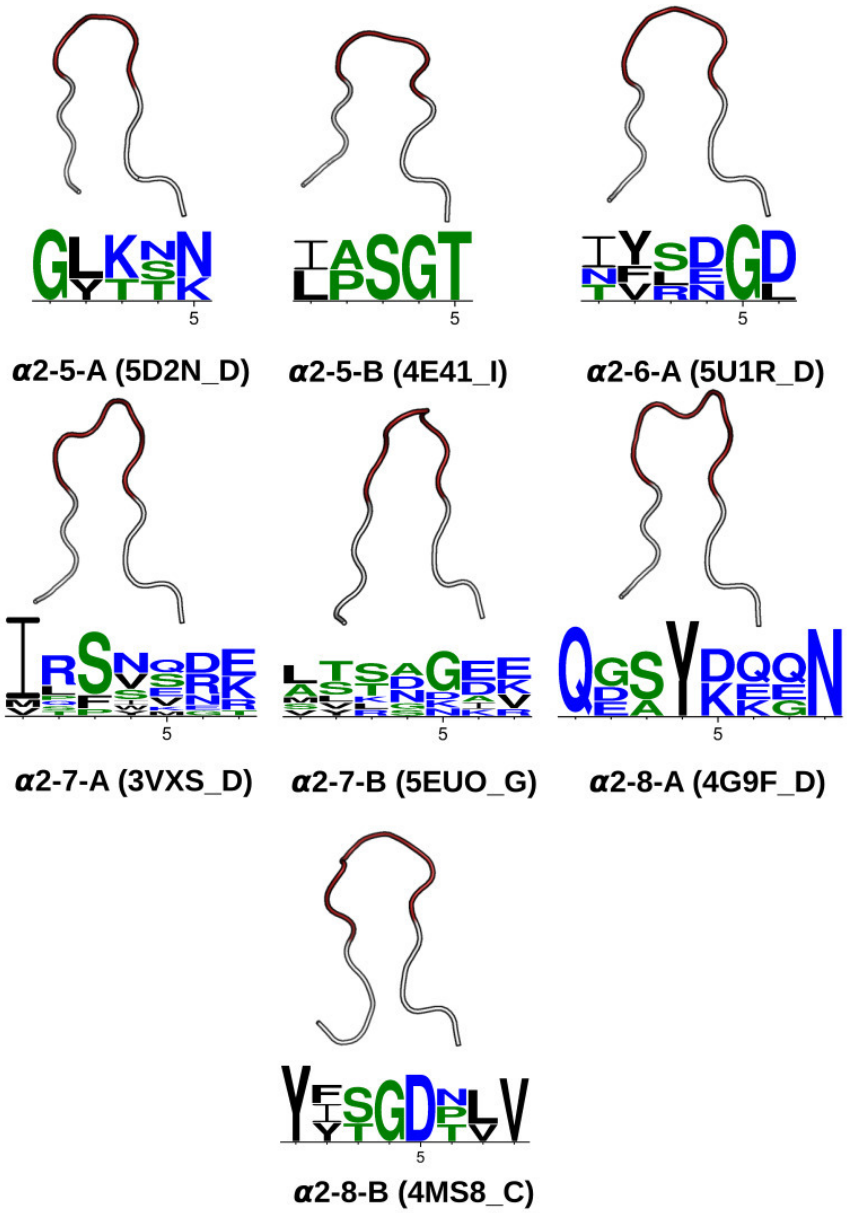


Figure B.2: CDRα2 loop clusters. At a 1.0Å clustering threshold, our DBSCAN method identified seven canonical classes (see Section 3.2.4). Each class has at least five structures and two unique sequences. Anchors are coloured white, while the CDRα2 region (IMGT 56–65) is coloured. The PDB four-letter code and the chain identifier of the centroid structure is shown in the bracket next to the cluster name. The sequence pattern below each centroid structure is generated by WebLogo (Crooks et al., 2004), using the unique sequences of the cluster: α2-5-A has 3, α2-5-B has 2, α2-6-A has 4, α2-7-A has 12, α2-7-B has 8, α2-8-A has 4 and α2-8-B has 3.

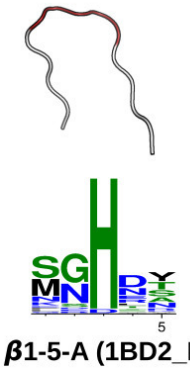


Figure B.3: CDRβ1 loop clusters. At a 1.0Å clustering threshold, our DBSCAN method identified one canonical class (see Section 3.2.4). Each class has at least five structures and two unique sequences. Anchors are coloured white, while the CDRβ1 region (IMGT 27–38) is coloured. The PDB four-letter code and the chain identifier of the centroid structure is shown in the bracket next to the cluster name. The sequence pattern below each centroid structure is generated by WebLogo (Crooks et al., 2004), using the unique sequences of the cluster: β1-5-A has 30.

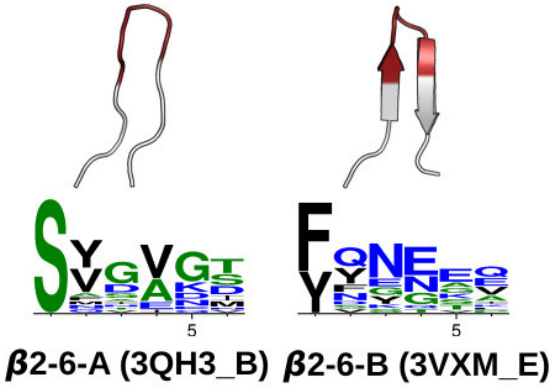


Figure B.4: CDRβ2 loop clusters. At a 1.0Å clustering threshold, our DBSCAN method identified two canonical classes (see Section 3.2.4). Each class has at least five structures and two unique sequences. Anchors are coloured white, while the CDRβ2 region (IMGT 56–65) is coloured. The PDB four-letter code and the chain identifier of the centroid structure is shown in the bracket next to the cluster name. The sequence pattern below each centroid structure is generated by WebLogo (Crooks et al., 2004), using the unique sequences of the cluster: β2-6-A has 18 and β2-6-B has 16.

B.2 Species information by cluster

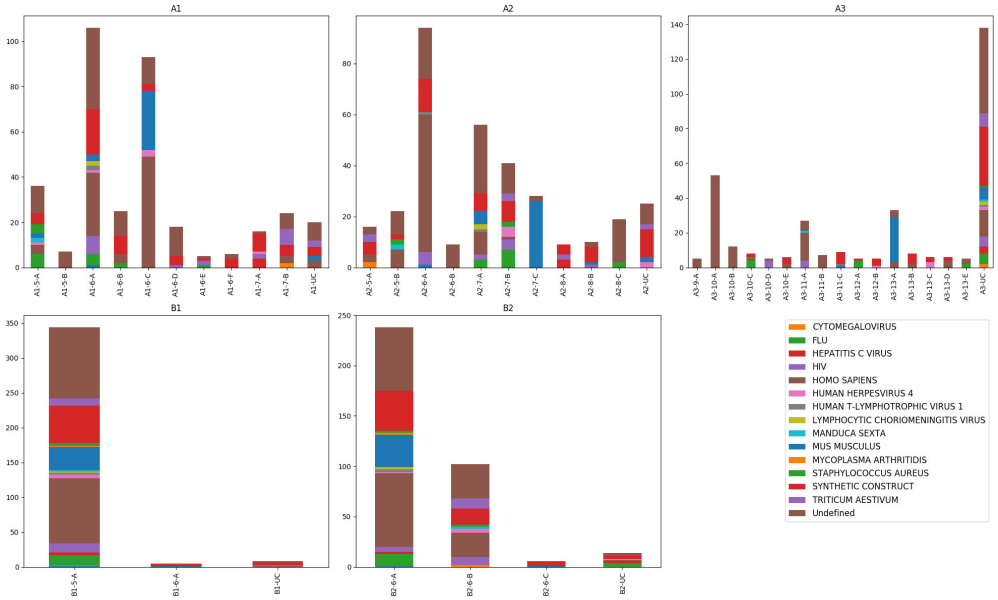


Figure B.5: Antigenic peptide organism information from the metadata of the TCR structures, sorted by CDR types and canonical forms.

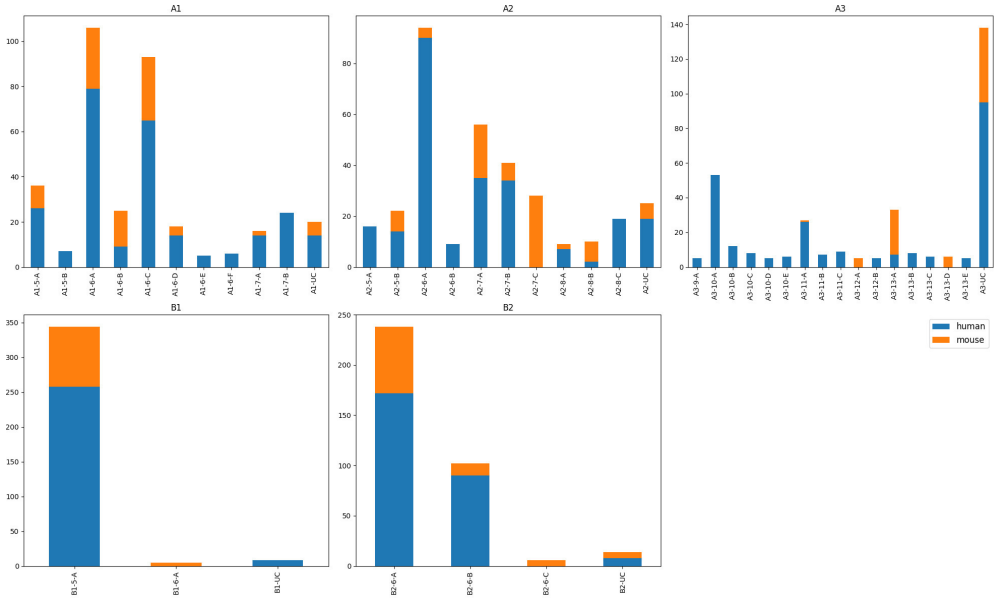


Figure B.6: Species information from the sequence alignment of the TCR structures, sorted by CDR types and canonical forms.

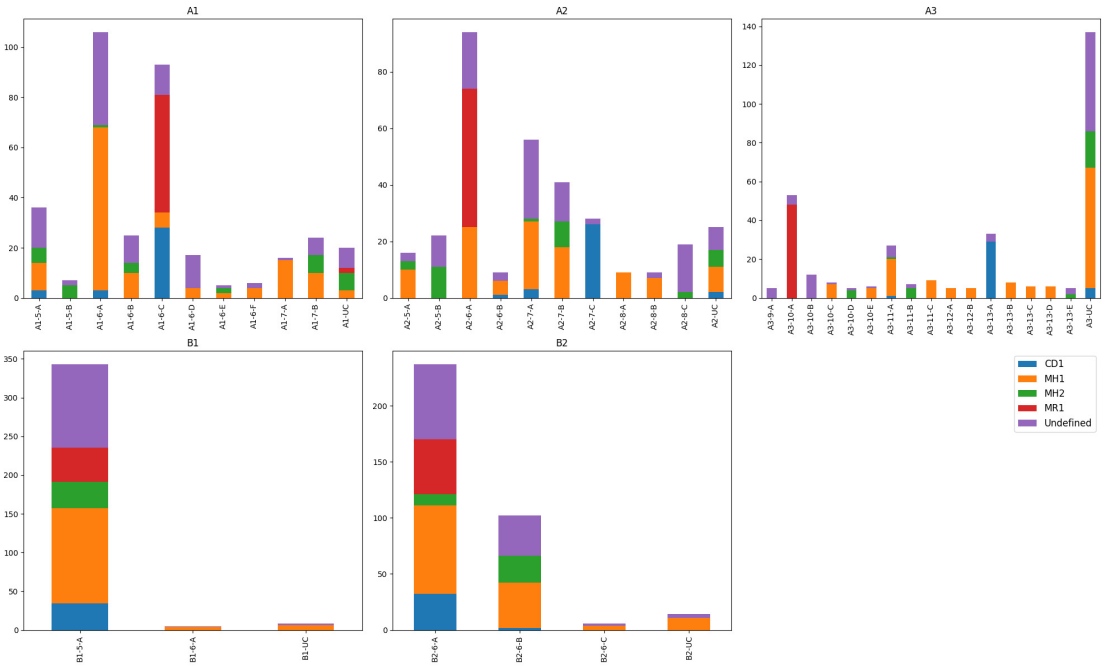


Figure B.7: Types of major histocompatibility complex (MHC) from the alignment of the MHC sequences in the corresponding TCR structures, sorted by CDR type and canonical forms. CD1: Cluster of differentiation 1; MH1: MHC Class I; MH2: MHC Class II and MR1: Major histocompatibility complex class I-related protein.

B.3 Prediction on Ig-seq dataset

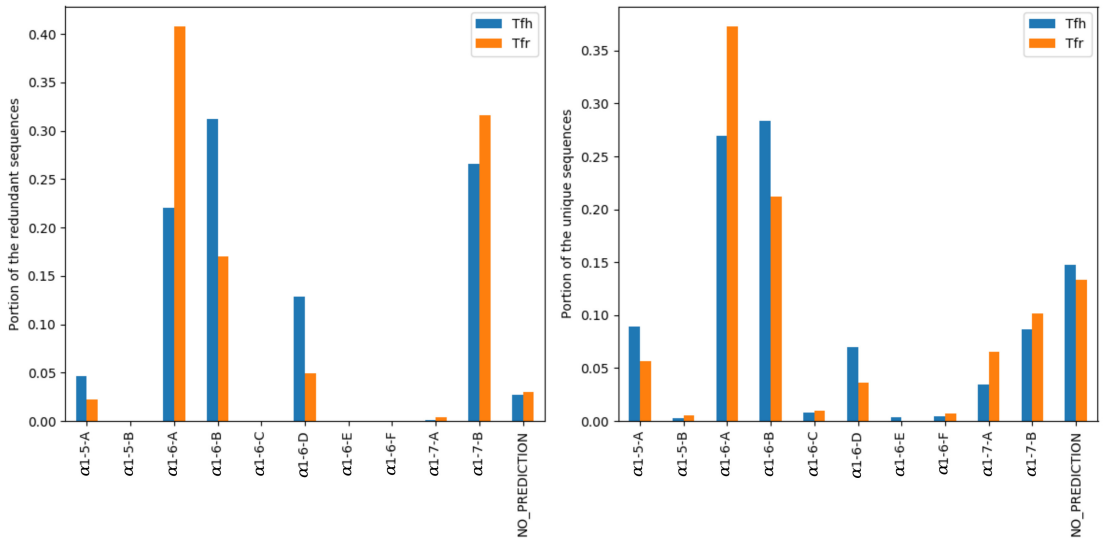


Figure B.8: CDRα1 canonical form prediction from the TCR sequences in Tfh and Tfr cells from Maceiras et al. (2017).

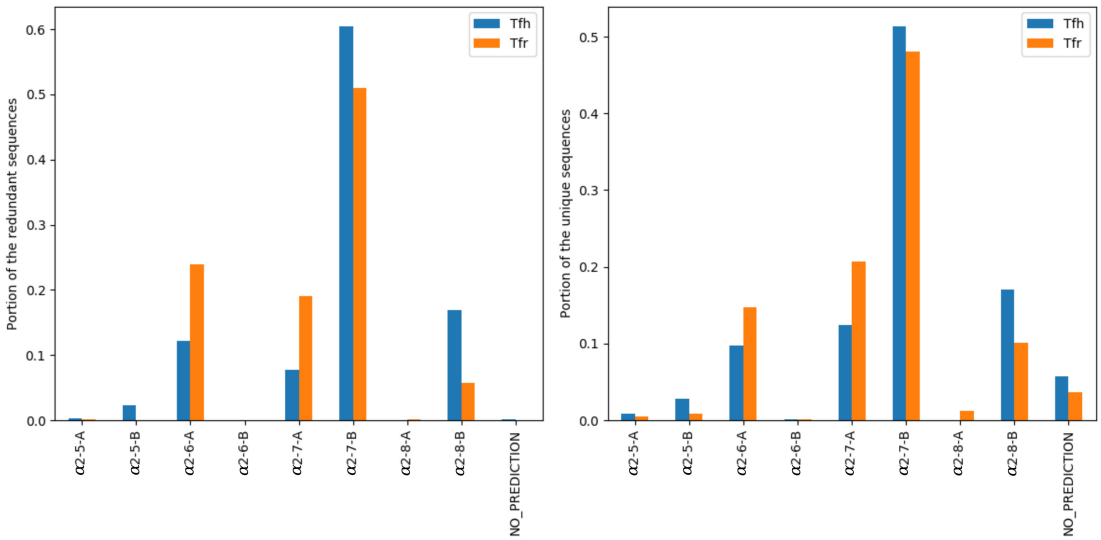


Figure B.9: CDRα2 canonical form prediction from the TCR sequences in Tfh and Tfr cells from Maceiras et al. (2017).

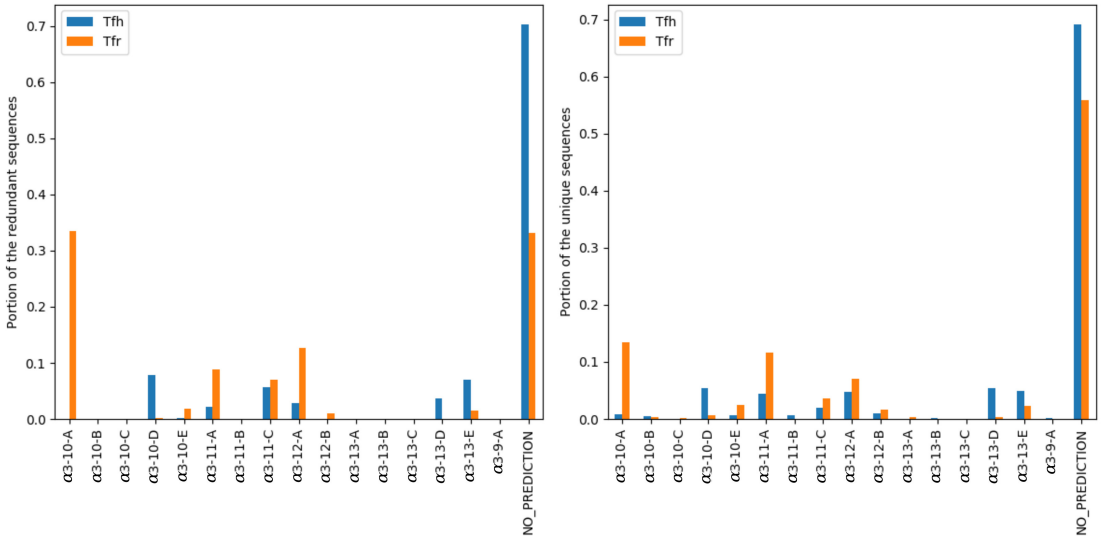


Figure B.10: CDRα3 canonical form prediction from the TCR sequences in Tfh and Tfr cells from Maceiras et al. (2017).

B.4 Comparison between TCR and antibody CDR structures

B.4.1 CDRα1/CDRL1

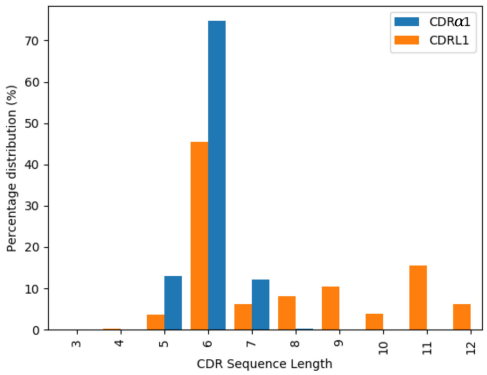


Figure B.11: Length distributions of CDRα1 (blue) and CDRL1 (orange) loops. The length distributions overlap between CDRα1 and CDRL1.

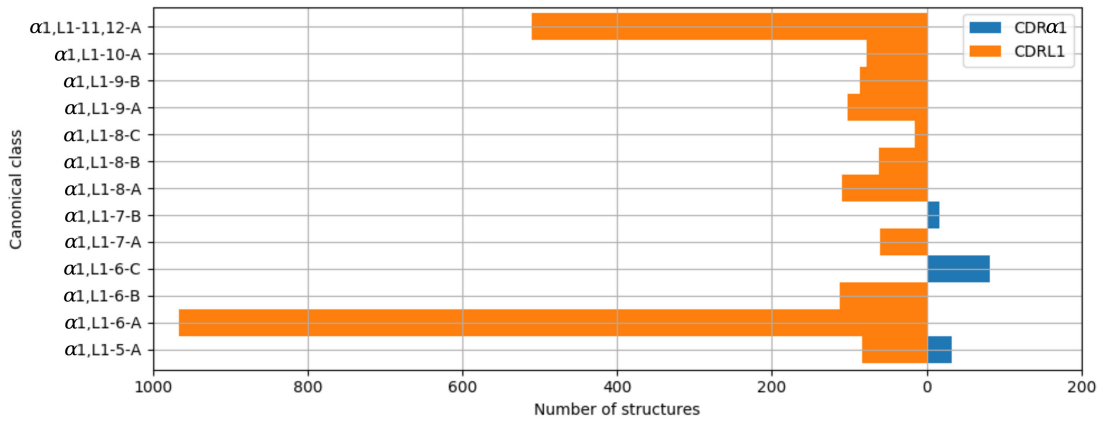


Figure B.12: Clusters of CDRα1 (blue) and CDRL1 (orange) loops. Orange bars indicate the number of CDRL1 structures, and blue bars the number of CDRα1 structures. All classes, apart from α1,L1-5-A, have structures from only one CDR type, *i.e.* CDRα1 or CDRL1.

B.4.2 CDRα2/CDRL2

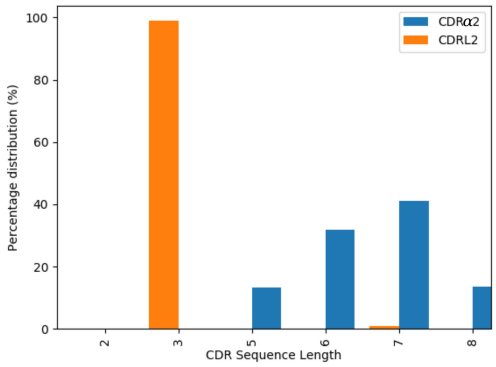


Figure B.13: Length distribution of CDRα2 (blue) and CDRL2 (orange) loops. The majority of the CDRL2 loops are three-residues long, while CDRα2 adopts a range of sequence lengths.

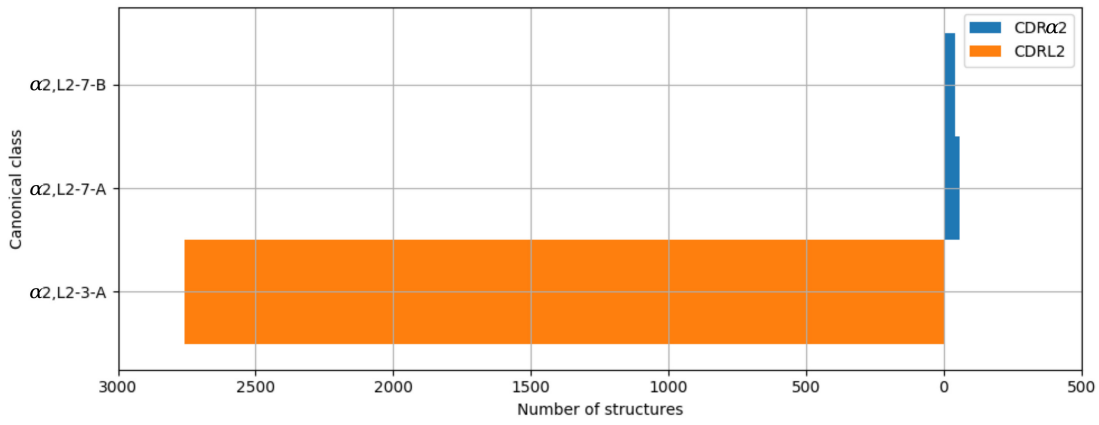


Figure B.14: Clusters of CDRα2 (blue) and CDRL2 (orange) loops. Orange bars indicate the number of CDRL2 structures, and blue bars the number of CDRα2 structures. All clusters contain structures from only one CDR type, *i.e.* CDRα2 or CDRL2.

B.4.3 CDRβ1/CDRH1

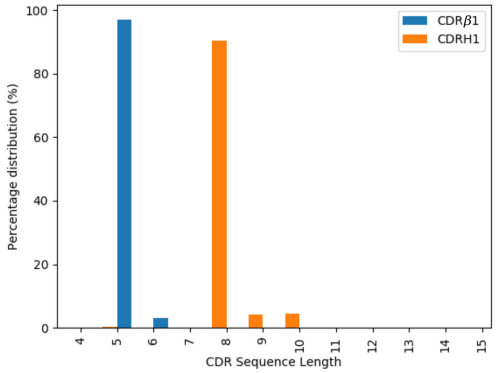


Figure B.15: Length distributions of CDRβ1 (blue) and CDRH1 (orange) loops. CDRβ1 loops tend to be shorter than CDRH1 loops.

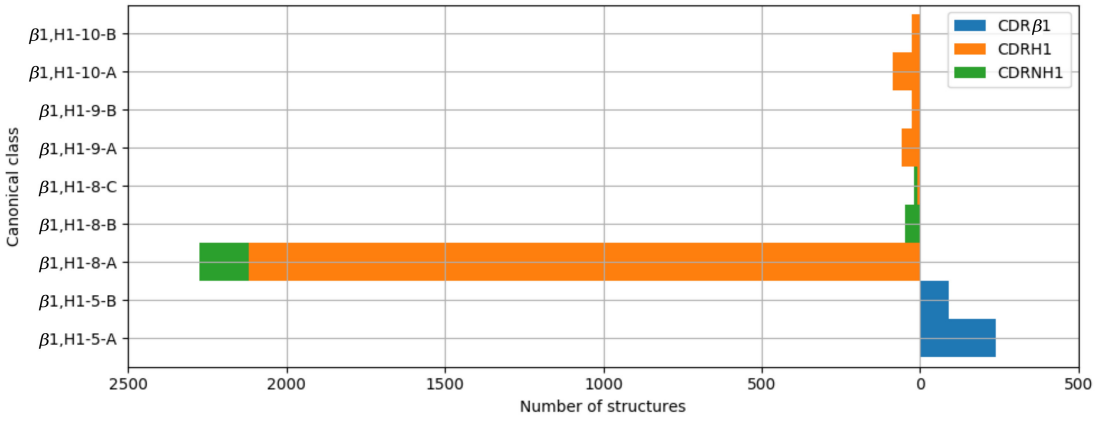


Figure B.16: Clusters of CDRβ1 and CDRH1 loops. Orange bars indicate the number of antibody CDRH1 structures, green bars nanobody CDRH1 structures, and blue bars the number of CDRβ1 structures. All classes have structures from only one CDR type, *i.e.* CDRβ1 or CDRH1. Some nanobody CDRH1 loops are structurally closer to antibody CDRH1 loops than TCR CDRβ1 loops.

B.4.4 CDRβ2/CDRH2

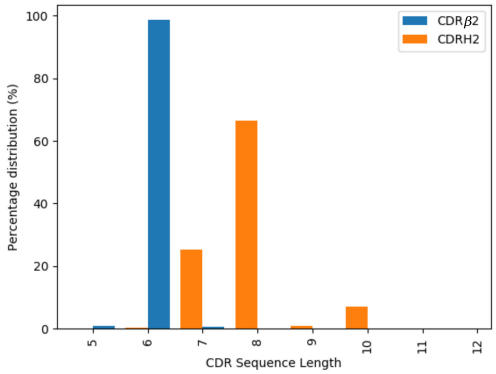


Figure B.17: Length distributions of CDRβ2 (blue) and CDRH2 (orange) loops. The majority of the CDRβ2 loops are slightly shorter than CDRH2 loops.

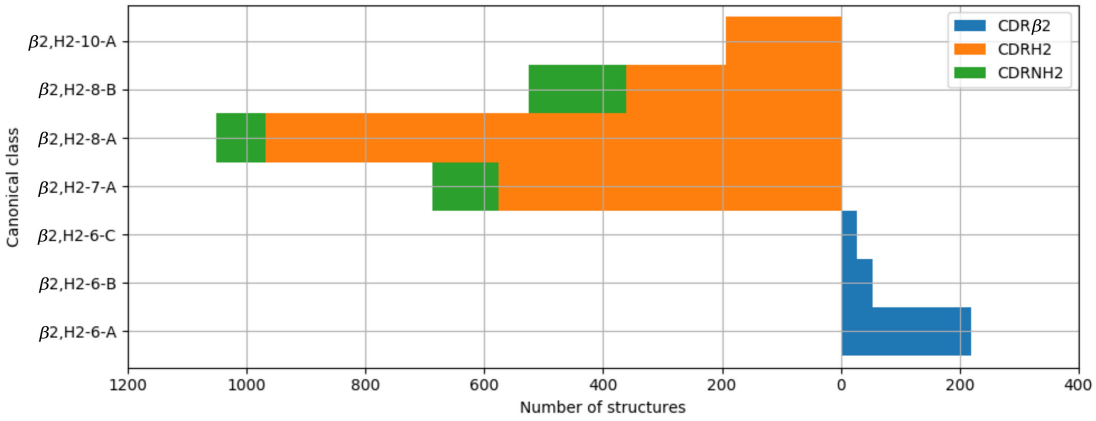


Figure B.18: Clusters of CDRβ2 and CDRH2 loops. Orange bars indicate the number of antibody CDRH2 structures, green bars nanobody CDRH2 structures, and blue bars the number of CDRβ2 structures. All classes have structures from only one CDR type, *i.e.* CDRβ2 or CDRH2. Antibody and nanobody CDRH2 loops are structurally more similar than TCR CDRβ2 loops.

B.4.5 CDRβ3/CDRH3

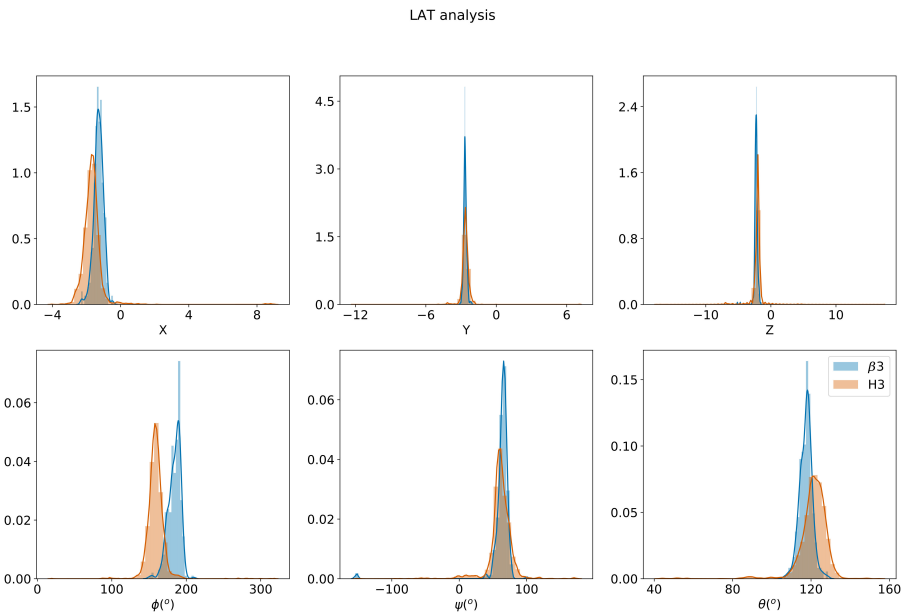


Figure B.19: Loop anchor transformation (LAT) analysis of CDRβ3 and CDRH3 structures. The Euler transformation was calculated on IMGT positions 105 and 117 (see Section 3.2.8). X, Y, Z, ϕ, ψ and θ are the six degrees of freedom in the Euler transformation.

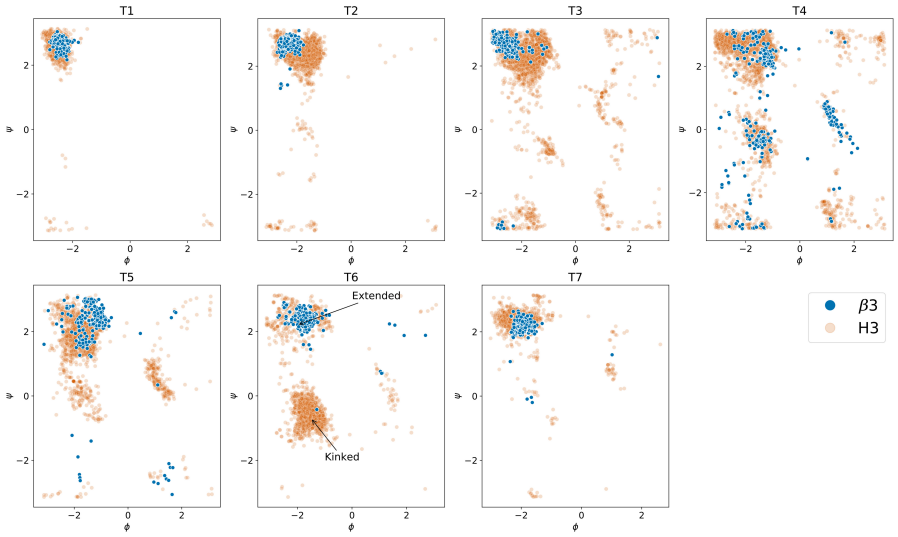


Figure B.20: The ϕ/ψ plot of the first three (T1-T3, corresponding to IMGT positions 105-107) and last four (T4-T7, corresponding to IMGT positions 114-117) loop residues, in CDRβ3 and CDRH3. Residue T6 (IMGT position 116) populates different regions of the plot depending on whether the torso is extended or kinked (as labelled). See Section 3.2.8 and Finn et al. (2016) for details.

B.5 CDR structural variability in TCR and antibodies

Table B.1: Number of sequences with multiple conformations.

	TCR and antibody (antibody subset)					
	L1	L2	L3	H1	H2	Total
# Unique sequences	580	198	855	793	885	3311
# Unique sequences with multiple structures	362	151	527	503	546	2089
# Unique sequences with structures in multiple classes	25	12	17	42	70	166
	TCR and antibody (TCR subset)					
	$\alpha 1$	$\alpha 2$	$\alpha 3$	$\beta 1$	$\beta 2$	Total
# Unique sequences	45	44	99	33	39	260
# Unique sequences with multiple structures	39	35	72	30	36	212
# Unique sequences with structures in multiple classes	11	6	9	3	12	41

Table B.2: TCR CDRs with multiple conformations that fall into different structural clusters (including canonical forms and pseudo-classes). 'B' and 'U' refer to the number of bound and unbound structures of the sequence within that cluster. NC: Non-clustered.

CDR	Sequence	Cluster 1	B	U	Cluster 2	B	U	Cluster 3	B	U
α1	ATGYPS	NC	4	-	α1,L1-6-H*	-	1	-	-	-
	DRGSQS	NC	-	2	α1,L1-6-C	16	5	-	-	-
	DRNFQY	NC	-	1	α1,L1-6-C	-	3	-	-	-
	DSAIYN	NC	1	-	α1,L1-6-J*	13	10	-	-	-
	DSSSTY	NC	1	-	α1,L1-6-C	1	-	-	-	-
	DSVNN	NC	1	-	α1,L1-5-A	3	4	α1,L1-5-B*	5	2
	YSATPY	NC	-	1	α1,L1-6-D*	4	-	α1,L1-6-E*	4	-
	VSGLRG	NC	3	-	α1,L1-6-G*	6	-	-	-	-
	TSGFNG	NC	6	4	α1,L1-6-G*	51	-	-	-	-
	TISGTDY	NC	-	1	α1,L1-7-E*	12	-	-	-	-
	SSVPPY	α1,L1-6-D*	2	5	α1,L1-6-E*	-	13	-	-	-
α2	QEAYKQQN	NC	1	-	α2,L2-8-A*	3	-	-	-	-
	ATKADDK	NC	4	-	α2,L2-7-B	2	1	-	-	-
	SSTDNKR	NC	-	2	α2,L2-7-B	-	2	-	-	-
	YYSGDPVV	NC	1	-	α2,L2-8-B*	7	1	-	-	-
	GLTSN	NC	3	1	α2,L2-5-A*	6	-	-	-	-
	GLKNN	NC	-	2	α2,L2-5-A*	4	2	-	-	-
α3	AMRGDSSYKLI	NC	-	1	α3,L3-11-B*	15	1	-	-	-
	AGAGSQGNLI	NC	1	1	α3,L3-10-H*	7	1	-	-	-
	AVNFGGGKLI	NC	-	2	α3,L3-10-J*	5	-	-	-	-
	GTYNQGGKLI	NC	2	-	α3,L3-10-A	1	-	-	-	-
	AYGEDDKII	NC	3	-	α3,L3-9-D*	2	-	-	-	-
	AVTTDSWGKLQ	NC	-	2	α3,L3-11-J*	9	-	-	-	-
	AVSESPFGNEKLT	NC	-	1	α3,L3-13-H*	2	3	-	-	-
	AVRPTSGGSYIPT	NC	-	1	α3,L3-13-C*	5	-	-	-	-
	AVRPLLDGTYIPT	NC	-	1	α3,L3-13-C*	3	-	-	-	-
β1	SEHNR	β1,H1-5-B	1	5	β1,H1-5-D*	2	1	-	-	-
	MGHDK	NC	1	-	β1,H1-5-A	7	6	-	-	-
	ENHRY	NC	2	-	β1,H1-5-A	4	1	-	-	-
β2	FQNEAQ	NC	1	-	β2,H2-6-C	8	9	-	-	-
	FNNNVP	NC	1	-	β2,H2-6-B	15	2	-	-	-
	FQNQEV	NC	-	1	β2,H2-6-D*	4	3	-	-	-
	FYNGKV	NC	-	1	β2,H2-6-B	-	3	-	-	-
	SASEGT	NC	2	-	β2,H2-6-A	47	-	-	-	-
	SFDVKD	β2,H2-6-A	-	6	β2,H2-6-F*	-	6	-	-	-
	SVGEGT	NC	-	2	β2,H2-6-A	3	3	-	-	-
	SYGAGN	NC	2	-	β2,H2-6-A	2	-	-	-	-
	SYGAGS	NC	1	-	β2,H2-6-A	34	10	-	-	-
	SYGVNS	NC	3	4	β2,H2-6-A	5	2	-	-	-
	FQGTGA	NC	1	-	β2,H2-6-C	2	-	-	-	-
	YYNGEE	NC	1	-	β2,H2-6-B	14	-	-	-	-

Table B.3: Antibody CDRs with multiple conformations that fall into different structural clusters (including canonical forms and pseudo-classes). 'B' and 'U' refer to the number of bound and unbound structures of the sequence within that cluster. NC: Non-clustered.

CDR	Sequence	Cluster 1	B	U	Cluster 2	B	U	Cluster 3	B	U
L1	QDISNY	NC	-	1	α1,L1-6-A	30	29	-	-	-
	QDINSY	NC	2	-	α1,L1-6-A	5	2	-	-	-
	KSVSTSGYSY	NC	1	-	α1,L1-10-A	1	-	-	-	-
	KSVSTSGYNY	NC	1	-	α1,L1-10-A	1	-	-	-	-
	ESVDYYGKSF	NC	-	2	α1,L1-10-A	2	-	-	-	-
	ENVVTY	NC	-	1	α1,L1-6-A	1	3	-	-	-
	QSVSSSY	α1,L1-7-A	6	2	α1,L1-7-C*	-	3	-	-	-
	QNLLDSSFDTNT	NC	-	1	α1,L1-11,12-A	-	1	-	-	-
	QSVSSY	NC	-	3	α1,L1-6-A	11	12	-	-	-
	QSVTNY	NC	1	1	α1,L1-6-I*	2	-	-	-	-
	QSVTSSQ	NC	-	1	α1,L1-7-A	-	1	-	-	-
	SRVGY	NC	-	1	α1,L1-5-A	-	3	-	-	-
	SSDVGGYNY	NC	22	4	α1,L1-9-A	15	13	-	-	-
	SSNIGENS	NC	-	1	α1,L1-8-A	-	5	-	-	-
	QGIRND	NC	1	-	α1,L1-6-A	2	-	-	-	-
	SSVSSSY	NC	2	-	α1,L1-7-A	4	4	-	-	-
	SSVSY	NC	2	-	α1,L1-5-A	23	16	-	-	-
	SSVTY	NC	-	1	α1,L1-5-A	-	7	-	-	-
	TGAVTSGHY	NC	1	-	α1,L1-9-B	7	-	-	-	-
	SSNVGGYNY	NC	-	1	α1,L1-9-A	-	1	-	-	-
	QSIVHSNGNTY	NC	-	1	α1,L1-11,12-A	30	17	-	-	-
	QSLSSKY	NC	-	1	α1,L1-7-A	-	1	-	-	-
	QSLVHSNGNTY	NC	1	1	α1,L1-11,12-A	26	16	-	-	-
	QSVDYDGDSY	NC	-	1	α1,L1-10-A	1	3	-	-	-
	QSVSSA	NC	11	-	α1,L1-6-A	22	-	α1,L1-6-I*	3	-
L2	LTS	NC	-	1	α2,L2-3-A	5	-	-	-	-
	GNN	NC	4	-	α2,L2-3-A	12	2	-	-	-
	EVN	NC	-	1	α2,L2-3-A	35	25	-	-	-
	GAS	NC	1	-	α2,L2-3-A	101	42	-	-	-
	FTS	α2,L2-3-A	5	1	α2,L2-3-D*	-	1	-	-	-
	GKN	α2,L2-3-E*	6	1	α2,L2-3-B*	2	5	-	-	-
	DVN	NC	2	-	α2,L2-3-A	2	3	-	-	-
	YTS	NC	-	1	α2,L2-3-A	64	37	α2,L2-3-D*	1	3
	DST	NC	1	-	α2,L2-3-A	-	1	-	-	-
	SDS	α2,L2-3-A	1	-	α2,L2-3-B*	1	-	-	-	-
	DAS	NC	-	1	α2,L2-3-A	86	78	-	-	-
	GDN	α2,L2-3-A	1	-	A2,L2-3-B*	1	-	-	-	-
L3	ALWYSNHLV	NC	1	1	α3,L3-9-B	13	8	-	-	-
	YSTDSSGNHRV	NC	1	-	α3,L3-11-A	-	2	-	-	-
	SSYGGDNNLF	NC	-	1	α3,L3-10-A	-	2	-	-	-
	SSYEGSDNFV	NC	17	4	α3,L3-10-C	16	5	-	-	-
	SSRDKSGSRLSV	NC	-	1	α3,L3-12-C*	5	-	-	-	-
	QVYGASSYT	NC	4	-	α3,L3-9-D*	-	6	-	-	-
	QQWSSHIFT	NC	-	1	α3,L3-9-B	-	1	α3,L3-9-C	-	1
	QQWNYPFT	NC	-	1	α3,L3-8-D*	2	1	-	-	-
	QAWDASTGV	NC	1	-	α3,L3-9-G*	7	-	-	-	-
	QQHYTTPPT	α3,L3-9-A	14	18	α3,L3-9-F*	-	6	-	-	-
	QAWDSSTVV	NC	-	1	α3,L3-9-B	1	2	-	-	-
	HQYLSSWT	NC	-	1	α3,L3-8-A	-	2	-	-	-
	GVGDTIKEQFVYV	NC	2	3	α3,L3-13-F*	5	-	-	-	-
	GTWDSSLNPV	NC	-	1	α3,L3-10-A	1	1	-	-	-
	AAWDDSLDVAV	NC	-	2	α3,L3-10,11-A	-	4	-	-	-
	QQSSSSLIT	α3,L3-9-C	12	-	α3,L3-9-A	1	-	-	-	-
	FQGSHVPRT	NC	1	1	α3,L3-9-A	2	6	-	-	-
H1	RRSSRSWA	NC	-	5	β1,H1-8-D*	-	2	β1,H1-8-T*	6	-
	GRTFSSYG	NC	1	-	β1,H1-8-C	3	-	-	-	-
	GRTFSSYA	NC	-	1	β1,H1-8-A	1	-	β1,H1-8-C	1	-
	GHTYSTYC	NC	1	2	β1,H1-8-F*	2	-	-	-	-
	GGTFSSYA	NC	3	4	β1,H1-8-A	2	2	-	-	-
	GGSISTYY	NC	1	-	β1,H1-8-A	1	-	-	-	-
	GGSISSSSYY	NC	-	1	β1,H1-10-B	-	1	-	-	-
	GGSFSTYA	NC	-	3	β1,H1-8-A	28	2	-	-	-

Continued on next page

Table B.3 – continued from previous page

CDR	Sequence	Cluster 1	B	U	Cluster 2	B	U	Cluster 3	B	U
	GRTFSSYV	β1,H1-8-A	-	1	β1,H1-8-C	7	-	-	-	-
	GGSFSSYY	NC	1	-	β1,H1-8-A	2	-	-	-	-
	GFTFSNYG	NC	1	-	β1,H1-8-A	15	4	-	-	-
	GFTFSDYD	NC	3	-	β1,H1-8-A	1	-	-	-	-
	GFSLSTSGIG	NC	1	-	β1,H1-10-A	1	-	-	-	-
	GFSLSDKA	NC	-	2	β1,H1-8-A	-	1	-	-	-
	GFSLRTSRVG	NC	1	2	β1,H1-10-C*	1	-	-	-	-
	GFSFNTNA	NC	1	1	β1,H1-8-A	2	2	-	-	-
	GFNIKDTY	NC	1	1	β1,H1-8-A	37	36	-	-	-
	GFTLDDYA	NC	1	-	β1,H1-8-A	2	-	-	-	-
	GSAVSDYA	NC	-	1	β1,H1-8-S*	-	5	-	-	-
	GSISGIVV	NC	-	1	β1,H1-8-B	3	-	-	-	-
	GSSFTGYN	NC	1	-	β1,H1-8-A	2	-	-	-	-
	GYTFTSYW	NC	4	-	β1,H1-8-A	34	30	-	-	-
	GYTFTSNW	β1,H1-8-A	1	6	β1,H1-8-C	-	1	-	-	-
	GYTFTSHW	NC	-	1	β1,H1-8-A	-	1	-	-	-
	GYTFTNYY	NC	1	1	β1,H1-8-A	4	1	-	-	-
	GYTFTNYG	NC	1	1	β1,H1-8-A	14	3	-	-	-
	GYTFTEYF	NC	-	1	β1,H1-8-A	3	-	-	-	-
	GYTFSEYW	NC	-	1	β1,H1-8-A	2	-	-	-	-
	GYSLSTSGMG	NC	-	1	β1,H1-10-A	-	1	-	-	-
	GYSITTNYA	β1,H1-9-D*	6	-	β1,H1-9-B	1	1	-	-	-
	GYSITSNYA	β1,H1-9-B	2	-	β1,H1-9-A	-	1	-	-	-
	GYSITSGYS	β1,H1-9-A	4	-	β1,H1-9-C*	3	2	-	-	-
	GYSITSDYA	NC	2	-	β1,H1-9-A	15	5	-	-	-
	GYSITSDFA	β1,H1-9-B	4	7	β1,H1-9-A	1	3	-	-	-
	GYSITGGYS	NC	-	2	β1,H1-9-A	-	6	-	-	-
	GYAFSSSW	NC	-	2	β1,H1-8-A	3	1	-	-	-
	GVTFSNVA	NC	1	-	β1,H1-8-A	1	-	-	-	-
	GVRLSAYD	NC	-	1	β1,H1-8-A	-	1	-	-	-
	GFNFSSSS	NC	1	-	β1,H1-8-A	2	-	-	-	-
	GDSITSGY	NC	1	-	β1,H1-8-A	17	2	-	-	-
	GSISSITT	NC	1	-	β1,H1-8-B	-	2	-	-	-
	GDSITSAY	NC	1	-	B1,H1-8-A	1	-	-	-	-
H2	ILPGSDST	NC	1	2	β2,H2-8-A	1	-	-	-	-
	IGPSGGIT	NC	-	4	β2,H2-8-B	-	4	-	-	-
	IDPSNGDT	NC	-	1	β2,H2-8-A	-	1	-	-	-
	IDPNSGGT	NC	2	-	β2,H2-8-A	11	8	-	-	-
	IDPNGGGT	NC	1	-	β2,H2-8-A	6	1	-	-	-
	IDPEQGNT	NC	-	2	β2,H2-8-A	1	1	-	-	-
	ISAGGDKT	NC	2	-	β2,H2-8-B	2	2	-	-	-
	IGSSGGQT	β2,H2-8-B	1	-	β2,H2-8-C*	-	1	-	-	-
	ISASGGST	β2,H2-8-B	-	4	β2,H2-8-G*	2	-	-	-	-
	ISGSGGNT	NC	1	-	β2,H2-8-B	-	1	-	-	-
	FNPSNGRT	NC	-	1	β2,H2-8-A	-	1	-	-	-
	ISGSGGST	NC	1	-	β2,H2-8-A	-	2	β2,H2-8-B	25	21
	ASNGGII	NC	-	1	β2,H2-7-A	-	1	-	-	-
	ISNLDGST	NC	2	1	β2,H2-8-A	-	1	-	-	-
	ISPGNGDI	NC	-	1	β2,H2-8-A	2	-	-	-	-
	ISPYSGVT	NC	1	-	β2,H2-8-A	1	1	-	-	-
	ISGGGRNI	β2,H2-8-B	-	1	β2,H2-8-L*	4	-	-	-	-
	IGTSGNI	NC	1	1	β2,H2-7-A	4	1	-	-	-
	IHYRGTT	NC	-	1	β2,H2-7-A	1	1	-	-	-
	IIPIFGTA	NC	3	2	β2,H2-8-A	4	6	-	-	-
	IRSKSINSAT	NC	1	-	β2,H2-10-A	-	3	-	-	-
	IRSGGGRT	NC	3	-	β2,H2-8-B	1	-	-	-	-
	IRNKPYNYET	NC	1	-	β2,H2-10-A	2	-	-	-	-
	INTHSGVP	NC	1	-	β2,H2-8-A	1	-	-	-	-
	INSNGGST	NC	1	-	β2,H2-8-B	6	1	-	-	-
	INPGSDYT	NC	2	-	β2,H2-8-A	2	1	-	-	-
	INPGNGYT	NC	-	2	β2,H2-8-A	2	1	-	-	-
	INLNGGRT	NC	-	1	β2,H2-8-A	-	1	-	-	-
	INIYTGEP	NC	1	-	β2,H2-8-A	2	1	-	-	-
	INIGATYA	NC	1	-	β2,H2-8-B	1	-	-	-	-
	ILPGSGST	NC	3	-	β2,H2-8-A	4	5	-	-	-
	ILPGSGRT	NC	2	-	β2,H2-8-A	1	1	-	-	-

Continued on next page

Table B.3 – continued from previous page

CDR	Sequence	Cluster 1	B	U	Cluster 2	B	U	Cluster 3	B	U
	ILPGGGSN	NC	-	2	β2,H2-8-A	-	2	-	-	-
	IKSKTDGGTT	NC	-	1	β2,H2-10-A	-	5	-	-	-
	IIPILGIA	NC	1	-	β2,H2-8-A	1	-	-	-	-
	ISRSGSVT	NC	1	-	β2,H2-8-B	2	1	-	-	-
	IRSSVIT	NC	1	-	β2,H2-7-B*	6	1	-	-	-
	ISSGGGNT	NC	-	2	β2,H2-8-B	1	-	-	-	-
	ISSPGTI	NC	-	3	β2,H2-7-A	-	1	-	-	-
	VIPLLTIT	NC	4	-	β2,H2-8-A	25	6	-	-	-
	TNPRNGGT	NC	1	-	β2,H2-8-A	4	2	-	-	-
	TIPLFGKT	NC	-	1	β2,H2-8-A	1	1	-	-	-
	NNPGNGYI	NC	-	1	β2,H2-8-A	-	1	-	-	-
	MSHEGDKT	NC	-	1	β2,H2-8-B	-	1	-	-	-
	MDSGGGGT	NC	9	1	β2,H2-8-B	-	1	-	-	-
	LTTTGTA	β2,H2-7-A	3	1	β2,H2-7-C*	1	4	-	-	-
	IYWDDVE	NC	-	1	β2,H2-7-A	1	-	-	-	-
	IYWDDDK	NC	1	-	β2,H2-7-A	8	7	-	-	-
	IYSYYGST	NC	2	-	β2,H2-8-A	2	-	-	-	-
	IYSSYSYT	NC	1	-	β2,H2-8-A	1	-	-	-	-
	ISSGGGRT	NC	1	1	β2,H2-8-B	1	-	-	-	-
	IYPYYGST	β2,H2-8-A	1	-	β2,H2-8-D*	3	-	-	-	-
	IYPYSGST	β2,H2-8-A	2	-	β2,H2-8-D*	1	-	-	-	-
	IYRSGSRM	NC	-	1	β2,H2-8-B	-	1	-	-	-
	IYPTNGYT	NC	1	-	β2,H2-8-A	16	28	-	-	-
	ISWSGGST	NC	-	1	β2,H2-8-B	1	1	-	-	-
	ITGPGEGWSV	NC	1	-	β2,H2-10-A	9	-	-	-	-
	ITPAGGYT	NC	3	-	β2,H2-8-J*	-	18	-	-	-
	VYDSGDT	NC	-	3	β2,H2-7-A	-	1	-	-	-
	IWAGGST	NC	2	-	β2,H2-7-A	-	1	-	-	-
	IWGDGIT	NC	-	1	β2,H2-7-A	1	-	-	-	-
	ITYSGTT	NC	1	1	β2,H2-7-A	1	-	-	-	-
	IWSGGST	NC	-	1	β2,H2-7-A	7	10	-	-	-
	IWWDDDN	NC	1	-	β2,H2-7-A	1	-	-	-	-
	IWYSGSNT	NC	-	2	β2,H2-8-C*	-	4	β2,H2-8-K*	-	9
	IYPGNGDT	NC	1	-	β2,H2-8-A	4	11	-	-	-
	IYPGNVHA	NC	-	1	β2,H2-8-A	-	2	-	-	-
	IYPGSGGT	NC	-	1	β2,H2-8-A	-	2	-	-	-
	IWPSGGNT	NC	1	-	β2,H2-8-B	1	1	-	-	-
	IRWNGGST	NC	1	-	β2,H2-8-B	2	-	-	-	-

C
Ab-Ligity Appendix

C.1 Ab-Ligity pipeline

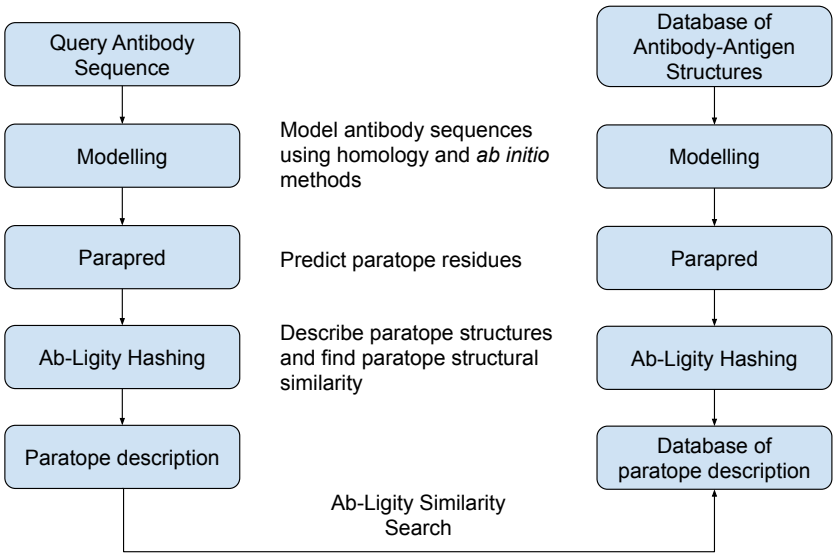


Figure C.1: Ab-Ligity pipeline. A database of antibody-antigen structures is modelled, by ABodyBuilder in this manuscript (Leem et al., 2016), has their paratopes predicted by Parapred (Liberis et al., 2018), and hashed by the Ab-Ligity algorithm described in the Methods section of the manuscript. This yields a database of paratope description in the form of Ab-Ligity hash tables. For a query antibody sequence, it would undergo homology modelling, paratope prediction and Ab-Ligity hashing to produce a paratope description. This query paratope description is then used to query a database of paratope description to find the Ab-Ligity similarity scores against known paratopes.

C.2 Performance evaluation

C.2.1 Selecting epitope similarity threshold

The epitope similarity score with the highest MCC is selected: 0.1 Ab-Ligity score for Ab-Ligity's definition of similar epitopes, and 0.7 InterComp score for that of InterComp.

Table C.1: Performance of the selected thresholds based on crystal paratope and crystal epitope similarities defined by the same method.

Methods	Paratope Similarity	Epitope Similarity	MCC	Precision	Recall
Ab-Ligity	0.1	0.1	0.94	0.98	0.89
InterComp	0.7	0.7	0.88	0.90	0.86

C.3 Parapred performance

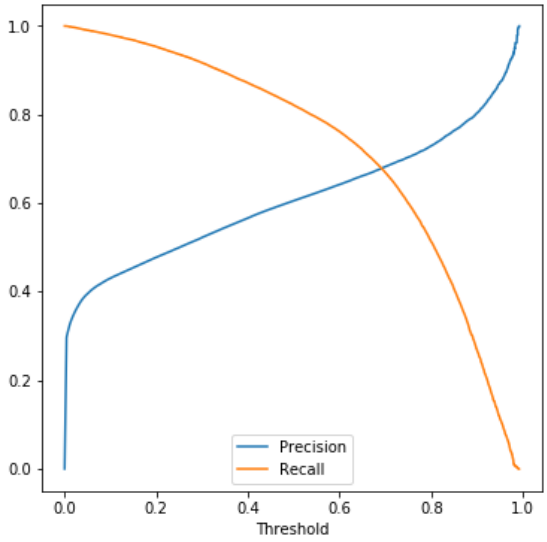


Figure C.2: Performance of Parapred across the threshold. Precision and recall are defined as in equations 2.4 and 2.5.

In this study, we used Parapred to predict the paratopes (Liberis et al., 2018). Parapred gives a score to each residue in the CDRs, and two residues before and two

after, to indicate how likely it will participate in binding. The precision and recall across different Parapred thresholds as calculated in the original paper (Liberis et al., 2018) are presented in Appendix Figure C.2.

In the original Parapred paper, they selected a threshold of 0.67 to balance between the actual and predicted paratope sizes. In the manuscript, we used this threshold to set the predicted paratopes for the Ab-Ligity and InterComp calculations.

D

Paratope analysis Appendix

D.1 Blind set data

Table D.1: PDB codes used in blind set evaluation. Only CDRs sequences with no exact matches in the training set and within the blind set are retained. 1 indicates the CDR in the PDB and chain code is used in the blind set; 0 means it is omitted. "|" in the Antigen Chain indicates that there are multiple non-antibody chains in the vicinity of the antibody, and the main antigen is a protein.

PDB ID	H Chain	L Chain	Antigen Chain	H1	H2	H3	L1	L2	L3
1A14	H	L	N	1	1	1	0	1	1
1A2Y	B	A	C	0	1	1	0	1	1
1AHW	B	A	C	1	1	1	1	0	1
1BJ1	H	L	W	0	1	1	0	1	1
1BQL	H	L	Y	1	0	1	1	0	1
1BVK	B	A	C	0	0	0	0	1	0
1DQJ	B	A	C	1	1	1	1	0	0
1DZB	A	A	X	0	0	1	1	0	0
1E6J	H	L	P	1	1	1	0	1	0
1FBI	H	L	X	0	1	1	0	0	1
1FE8	H	L	A	1	1	1	1	0	1
1FJ1	B	A	F	1	1	1	0	0	1
1G9M	H	L	G	1	1	1	1	0	1
1IQD	B	A	C	1	1	1	1	0	1
1JHL	H	L	A	1	1	1	1	0	1
1MLC	B	A	E	1	1	1	0	1	1
1N8Z	B	A	C	0	1	1	1	0	1
1NDG	B	A	C	1	1	1	0	0	0
1P2C	B	A	C	1	1	1	0	1	1
1PKQ	B	A	E	1	1	1	1	0	1
1QFU	H	L	A	1	1	1	1	0	1
1UJ3	B	A	C	0	1	1	1	0	0
1XF5	B	A	P | L	1	1	1	1	0	1
1YY9	D	C	A	1	1	1	1	0	1
2CMR	H	L	A	1	0	1	1	0	1
2EIZ	B	A	C	1	0	1	0	1	1
2GHW	B	B	A	1	1	1	1	0	1
2J6E	H	L	A	1	1	1	1	1	1
2NR6	D	C	A	0	1	1	1	0	1
2R0L	H	L	A	1	1	1	0	0	0

Continued on next page

Table D.1 – continued from previous page

PDB ID	H Chain	L Chain	Antigen Chain	H1	H2	H3	L1	L2	L3
2R29	H	L	A	0	1	1	1	1	1
2W9E	H	L	A	1	1	1	0	0	1
2XRA	H	L	A	1	1	1	1	0	1
2XTJ	D	B	A	1	1	1	1	1	1
3BGF	B	C	A	1	1	1	0	0	1
3BN9	D	C	B	0	0	1	1	0	1
3DVG	B	A	Y | X	1	1	1	0	1	1
3EOA	B	A	J	1	1	1	1	0	0
3G04	B	A	C	1	1	1	1	1	1
3HI1	B	A	J	1	1	1	1	0	1
3HMX	H	L	A	1	1	1	1	0	1
3I50	H	L	E	1	1	1	0	1	1
3L5W	B	A	J	1	0	0	0	0	1
3NFP	A	B	K	1	1	1	1	0	1
3P0Y	H	L	A	1	1	1	1	0	1
3S37	H	L	X	0	0	1	1	0	1
3SO3	C	B	A	0	0	1	0	0	1
3SQO	H	L	A	1	1	1	1	0	1
3U7Y	H	L	G	1	0	1	1	0	0
3UX9	B	B	A	0	0	1	0	0	1
3UYP	A	A	B	1	1	1	1	0	1
3X3F	H	L	A	1	1	1	0	0	1
4BZ2	H	L	A	1	1	1	1	0	1
4DKE	H	L	A	1	1	1	0	0	1
4DN4	H	L	M	1	0	1	1	0	1
4EDX	B	A	W	1	1	1	1	1	1
4F15	B	C	A	1	0	1	0	0	1
4F2M	A	B	E	0	1	1	0	0	1
4FP8	H	L	A	1	1	1	1	0	1
4G3Y	H	L	C	1	1	1	1	0	1
4G6J	H	L	A	1	1	1	1	0	1
4G6M	H	L	A	0	1	1	1	1	1
4GMS	H	L	A	0	1	1	1	0	1
4H8W	H	L	G	1	1	1	1	1	1
4HC1	H	L	A	0	1	1	1	1	1
4HJ0	C	D	B	0	1	1	1	0	1
4I77	H	L	Z	1	1	1	1	0	1
4JB9	H	L	G	1	1	1	1	0	0
4JPW	H	L	G	1	1	1	1	1	1
4L5F	H	L	E	1	1	1	0	1	1
4LU5	H	L	B	0	1	1	1	0	1
4M1G	H	L	B | A	1	1	1	1	0	1
4M5Z	H	L	A	1	1	1	1	0	1
4NIK	B	B	A	1	1	1	1	1	1
4O9H	H	L	A	1	1	1	1	1	1
4OII	H	L	A	0	1	1	0	0	0
4OLU	H	L	G	1	1	0	0	0	0
4PP1	D	C	B	0	1	1	1	0	1
4PP2	B	A	F	1	1	1	0	0	1
4PY8	I	J	B | A	1	1	1	0	0	1
4QHU	B	A	D	0	1	1	1	1	1
4QTI	H	L	U	1	1	1	1	0	1
4RGM	C	B	A	1	1	1	1	1	1
4S1Q	H	L	G	1	1	1	0	0	0
4TSA	H	L	A	1	1	1	1	1	0
4UTA	H	L	B	1	1	1	1	0	1
4UV7	H	L	A	0	0	0	1	0	1
4XAK	D	E	B	0	1	1	0	0	1
4XNM	B	A	D	1	1	1	1	1	1
4XNY	H	L	G	1	0	1	1	1	1
4XWG	H	L	A	0	1	1	1	0	1
4Y5V	A	B	C	1	0	1	1	1	1
4YDK	H	L	G	1	1	1	1	1	1
4YDL	B	C	A	1	1	1	1	1	1
4YJZ	L	L	E	1	1	1	1	1	1
4YK4	C	B	A	1	0	1	0	0	1

Continued on next page

Table D.1 – continued from previous page

PDB ID	H Chain	L Chain	Antigen Chain	H1	H2	H3	L1	L2	L3
4YPG	B	A	C	1	1	1	1	0	1
4Z5R	B	A	N	1	1	1	1	1	1
4ZPT	A	B	R	0	1	1	1	0	1
5BJZ	C	L	A	1	1	1	0	0	1
5BK1	C	D	B	1	0	1	0	0	1
5BK2	C	D	B	1	1	1	0	0	1
5C0N	C	D	B	1	1	1	1	0	1
5C7X	H	L	A	0	1	0	0	0	1
5CBA	A	B	E	0	0	1	1	1	1
5D8J	H	L	A	1	1	1	1	1	1
5D93	C	B	A	0	1	1	0	0	1
5DUR	B	D	A	1	1	1	1	1	1
5EII	A	B	I	1	1	1	0	0	1
5F3H	A	B	J	0	1	1	0	0	0
5F72	S	S	K	1	1	1	0	0	1
5F9O	H	L	G	1	1	0	1	0	1
5F9W	B	C	A	1	1	1	0	1	1
5GMQ	B	C	A	0	1	1	0	0	1
5GRU	H	H	A	0	1	1	1	0	1
5GRU	L	L	A	0	0	1	0	0	1
5HDQ	H	L	A	0	1	1	1	0	1
5HYS	A	B	G	1	1	1	0	0	1
5I9Q	B	C	A	1	1	1	1	1	1
5IES	H	L	C	0	0	1	1	0	1
5IWL	A	A	D	1	1	1	1	0	0
5JHL	H	L	A	1	1	1	1	1	1
5JYL	B	B	A	1	1	1	0	0	1
5JYM	B	B	A	1	1	1	1	0	1
5K9K	A	B	I	1	1	1	1	0	1
5KJR	H	L	G	1	1	1	1	0	1
5KVD	H	L	E	1	1	1	1	0	0
5KVE	H	L	E	1	1	1	1	1	0
5KVF	H	L	E	1	1	1	1	0	1
5KVG	H	L	E	1	1	1	1	0	1
5L6Y	H	L	C	1	1	1	1	1	1
5LDN	H	L	A	1	1	1	1	1	1
5LSP	H	L	A	1	1	1	0	0	1
5MHR	I	H	B	0	1	1	0	1	1
5MO9	H	L	X	1	1	1	1	0	1
5N7W	A	B	Y	1	0	1	1	0	1
5NGV	H	L	A	1	1	1	1	1	1
5NMV	H	L	K	0	1	1	1	0	1
5NUZ	A	B	C	0	1	1	1	1	1
5O14	C	D	B	1	1	1	1	1	1
5O1R	H	L	A	1	1	1	0	1	1
5O4G	B	A	C	1	1	1	0	0	1
5OB5	H	L	A	1	1	1	1	1	1
5OCC	H	L	A	1	1	1	0	1	1
5OGI	B	B	A	0	0	0	0	0	0
5OTJ	B	A	D	1	1	1	1	1	1
5SX4	H	L	N	1	1	1	0	0	1
5SY8	H	L	O	0	0	0	0	0	0
5T5F	H	L	A	1	1	1	1	1	1
5TLK	D	C	X	1	1	1	1	1	0
5TRU	H	L	C	1	1	1	1	1	1
5TUD	C	B	A	0	1	1	0	0	1
5TZT	A	B	D	0	1	1	1	0	0
5TZU	H	L	C	1	1	1	1	1	1
5UCB	H	L	B	1	1	1	0	0	1
5UGY	H	L	A	1	1	1	1	1	1
5USH	D	E	A	1	1	1	0	1	1
5USI	A	B	Y	0	1	1	1	1	1
5USL	B	C	A	1	1	1	1	1	1
5UTZ	B	C	A	1	0	1	1	1	1
5VAG	C	B	A	0	0	1	0	0	1
5VIC	H	L	E	1	1	1	1	1	1

Continued on next page

Table D.1 – continued from previous page

PDB ID	H Chain	L Chain	Antigen Chain	H1	H2	H3	L1	L2	L3
5VIG	A	B	G	1	1	1	1	1	1
5VKD	H	L	A	0	1	1	0	0	0
5VLP	H	L	A	1	1	1	1	0	1
5VYF	B	A	C	0	1	1	0	0	1
5W06	H	L	T	1	1	1	1	0	0
5W08	G	H	A	1	1	1	1	1	1
5W0D	B	C	A	0	0	1	1	1	1
5W5X	H	L	A	1	1	1	1	1	1
5WB9	H	L	G	1	1	1	1	1	1
5WHK	H	L	A | B	1	1	1	1	1	1
5WI9	F	E	B	0	1	1	0	0	1
5WK3	Q	P	A	0	1	1	1	0	1
5WT9	H	L	G	1	1	1	0	0	1
5X0T	A	B	E	1	1	1	1	0	1
5X8M	B	C	A	1	1	1	1	0	1
5XKU	C	B	A	0	0	1	0	0	1
5XMH	D	C	B	1	1	1	0	0	1
5XWD	H	D	A	1	0	1	0	1	0
5XXY	H	L	A	1	1	1	0	0	1
5Y11	A	B	C	1	1	1	0	0	1
5Y2L	I	J	B | A	1	1	1	0	1	1
5Y9J	H	L	A	1	1	1	0	0	1
5YOY	G	D	B	1	1	1	1	0	1
5YY5	C	D	B	1	1	1	1	1	1
6A0Z	H	L	A	1	1	1	1	0	1
6A3W	A	B	C	1	1	1	1	1	1
6A67	C	D	B	1	1	1	1	0	1
6A77	H	L	A	1	1	1	1	1	1
6AL0	H	L	A	1	1	1	1	1	1
6AL5	H	L	A	1	1	1	1	0	1
6AOD	B	A	C	0	1	1	0	0	1
6APB	H	L	C | A	1	1	1	0	0	1
6AQ7	H	L	A	1	1	1	1	1	1
6AZZ	C	B	A	1	1	1	1	1	1
6B08	C	B	A	1	0	1	0	1	1
6B0A	H	L	A	1	1	1	1	1	1
6B0G	D	C	E	0	1	1	1	0	1
6B0H	B	A	J	1	1	1	1	0	1
6B0S	H	L	C	0	1	1	1	0	1
6BCK	H	L	G	1	1	1	1	1	1
6BF4	B	C	A	1	1	1	1	1	1
6BFQ	A	B	I	1	1	1	1	0	1
6BFS	H	L	C	1	1	1	1	0	1
6BIT	I	K	H	1	1	1	1	0	1
6BKB	H	L	E	1	1	1	1	0	0
6BKC	H	L	E	1	1	1	0	0	1
6BPA	B	C	A	1	0	1	1	0	1
6BPC	B	C	A	1	1	1	0	0	0
6C6Z	C	D	A	1	0	1	1	1	1
6C9U	H	L	A	1	1	1	0	0	1
6CBV	H	L	B	1	1	1	0	0	1
6CMG	C	B	A	1	1	1	1	0	1
6CRK	H	H	A | B | G	1	1	1	1	0	1
6CVK	A	A	C	1	1	1	1	1	1
6CW2	A	B	D | C	0	0	1	0	0	1
6CW3	A	B	H | G	0	1	1	0	0	1
6CXY	H	L	C	1	1	1	1	1	1
6CYF	D	C	A	1	1	1	0	0	1
6DDM	B	A	C	1	1	1	1	0	1
6DDR	B	A	C	0	1	1	1	1	1
6DDV	B	A	C	1	1	1	1	0	0
6DFI	H	L	E	1	1	1	1	1	1
6DKJ	A	B	D	1	1	1	0	1	1
6E3H	H	L	B | A	1	1	1	1	1	1
6E4X	Z	Y	B	1	1	1	1	0	1
6E56	G	I	D	1	1	1	1	0	1

Continued on next page

Table D.1 – continued from previous page

PDB ID	H Chain	L Chain	Antigen Chain	H1	H2	H3	L1	L2	L3
6E62	B	C	A	1	1	1	1	1	1
6E63	B	C	A	0	1	0	0	0	1
6EJM	H	H	B	1	1	1	1	0	1
6EK2	H	H	A	1	1	1	0	0	1
6ELU	B	C	A	1	1	1	1	0	1
6EWB	E	F	A	1	1	1	1	1	0
6FAX	H	L	R	1	1	1	1	0	1
6FGB	H	L	A | B	1	1	1	1	1	1
6FLA	A	B	G	1	1	1	1	0	1
6FLC	B	A	I	1	1	1	1	0	1
6GKU	H	L	A	0	1	1	1	1	1
6GLW	C	D	B	1	1	1	1	0	1
6H2Y	H	L	D	1	1	1	1	1	1
6H3T	H	L	A	1	1	1	0	0	1
6HHC	H	L	A	1	1	1	0	0	1
6HIG	H	L	B	1	1	1	1	1	1
6HX4	H	L	B	0	1	1	1	0	1
6HXW	C	D	B	1	1	1	1	0	1
6I07	A	A	C	0	1	1	1	1	1
6I8S	E	I	A	0	0	1	1	0	0
6I9I	A	B	C	1	1	1	1	0	1
6IAP	E	D	A	1	1	1	1	0	1
6IAP	H	L	A	1	1	1	0	0	1
6ID4	C	D	E | F	0	1	1	1	1	1
6IEA	H	L	A	0	1	1	1	1	1
6IEB	E	F	B	1	0	1	1	1	1
6IEK	B	C	A	1	1	1	1	0	0
6ION	H	L	A	1	1	1	1	0	1
6IUT	H	L	A	1	1	1	1	1	1
6IUV	C	D	B	1	1	1	1	1	1
6IVZ	H	L	A	1	1	1	1	0	1
6J11	D	E	C	1	0	1	1	1	1
6J14	A	B	G	1	1	1	1	0	0
6J15	A	B	C	1	1	1	1	0	0
6J5D	H	L	A	1	1	1	0	1	1
6J6Y	B	C	A	1	1	1	1	1	1
6J71	B	B	A	1	1	1	1	0	1
6JBT	H	L	F	0	1	1	0	0	0
6JEP	H	L	E	1	1	1	1	0	1
6JJP	A	B	C	1	1	1	1	0	1
6K0Y	A	B	C	1	1	1	1	0	1
6K65	H	L	A	1	1	1	1	0	1
6MEH	H	L	C	1	1	1	0	0	1
6MEI	H	L	C	1	1	1	1	0	1
6MEJ	A	B	C	1	1	1	0	1	1
6MFP	C	D	A	1	1	1	1	1	1
6MG7	H	L	G	1	1	1	1	0	0
6MHR	A	B	C	1	1	1	0	0	1
6MLK	H	L	A	0	1	1	0	0	1
6MTO	H	L	T	1	1	1	1	0	1
6MTQ	H	L	T	1	1	1	1	0	1
6MVL	H	L	A	1	1	1	0	0	1
6N5D	C	D	A	0	1	1	0	1	0
6N6B	K	L	A	1	1	1	1	0	1
6N81	C	D	B	1	1	1	0	0	1
6NHA	H	L	A | B	0	1	1	1	0	1
6NMR	A	B	E	1	1	1	1	1	1
6NMS	B	A	C	1	1	1	1	0	1
6NMT	B	A	C	1	1	1	1	1	1
6NMU	B	A	C	1	1	1	1	1	1
6NMV	H	L	S	1	1	1	1	1	1
6NYQ	H	L	C	1	1	1	1	0	1
6NZ7	G	I	F | E	1	0	1	1	1	1
6O1F	H	L	A | I	1	1	1	1	1	1
6O39	B	A	C	1	1	1	0	0	1
6O3A	B	A	E	0	1	1	0	0	1

Continued on next page

Table D.1 – continued from previous page

PDB ID	H Chain	L Chain	Antigen Chain	H1	H2	H3	L1	L2	L3
6O3B	B	A	C	1	1	1	0	0	1
6O9H	A	B	C	1	1	1	0	0	1
6O9I	D	E	C	1	1	1	1	0	1
6OAN	B	B	A	1	1	1	1	1	1
6OC3	A	B	F	1	1	1	1	0	1
6OGX	C	D	G	1	1	1	0	0	1
6OGX	H	L	G	0	1	1	1	0	1
6OOR	H	L	A	1	1	1	1	1	1
6OTC	H	L	A	1	1	1	0	0	1
6OY4	D	C	A	0	1	1	1	0	0
6P50	H	L	C	0	1	1	1	0	1
6P67	A	B	K	1	1	1	1	0	1
6PCU	B	B	A	1	1	1	1	1	1
6PE8	A	B	U	1	1	1	1	0	0
6PHB	B	A	E	1	1	1	0	0	1
6PHC	A	B	I	1	1	1	1	0	1
6PI7	C	B	A	1	0	1	0	0	1
6PIS	H	L	B	1	1	1	1	0	1
6PLK	H	L	E	0	1	1	1	1	1
6PPG	B	A	G	0	1	1	0	0	1
6PXH	C	D	A	1	1	1	0	1	1
6PZE	H	L	A	0	0	1	1	0	1
6PZF	F	E	B	0	0	1	1	0	1
6Q0E	H	L	A	0	0	1	0	0	1
6Q0O	Z	B	A	1	1	1	1	0	1
6Q18	H	L	A	1	1	1	1	0	1
6Q20	H	L	A	1	1	1	1	1	1
6QIG	H	L	A	0	0	1	1	1	1
6R8X	C	B	A	1	1	1	1	1	1
6RPS	H	L	A	1	1	1	1	0	1
6S5A	H	L	D | A	1	1	1	0	0	0
6SVL	A	B	C	0	1	1	0	0	1
6TYB	H	L	G	1	1	1	1	1	1
6U36	H	L	B	1	1	1	1	0	1

D.2 Peptide-binding antibody set

Table D.2: PDB codes used in the peptide-binding antibodies evaluation. "|" in the Antigen Chain indicates that there are multiple non-antibody chains in the vicinity of the antibody, and the main antigen is a peptide.

PDB ID	H Chain	L Chain	Antigen Chain	PDB ID	H Chain	L Chain	Antigen Chain
1A3R	H	L	P	3QNZ	B	A	C
1ACY	H	L	P	3SGE	H	L	K
1BOG	B	A	C	3U0T	B	A	F
1CE1	H	L	P	3UJI	H	L	P
1CU4	H	L	P	3UJJ	H	L	P
1E4W	H	L	P	3UO1	H	L	P
1EJO	H	L	P	4DGV	H	L	A
1F58	H	L	P	4G6A	C	D	B
1F90	H	L	E	4G6F	H	L	P
1FPT	H	L	P	4H0H	B	B	D
1FRG	H	L	P	4HHA	B	A	P
1GGI	H	L	P	4HIX	H	L	A
1HIM	L	H	P	4HPO	H	L	P
1I8I	B	A	C	4HPY	H	L	P
1JP5	A	A	C	4HS6	B	A	Z
1KC5	H	L	P	4HS8	H	L	A
1KCR	H	L	P	4HZL	A	B	F
1KCS	H	L	P	4J8R	B	A	I
1KTR	H	L	P	4JFX	H	L	P
1MPA	H	L	P	4JO1	H	L	P
1MVU	B	A	P	4JO3	H	L	P
1N0X	H	L	P	4JZN	I	P	K
1N64	H	L	P	4JZO	G	H	L
1NAK	H	L	P	4LKX	A	B	R
1NL0	H	L	G	4N0Y	H	L	A
1P4B	H	L	P	4N8C	H	L	X
1PZ5	B	A	C	4N9G	A	B	C
1Q1J	H	L	P	4NRX	A	B	C
1QKZ	H	L	P	4O4Y	H	L	A
1SM3	H	L	P	4ONF	H	L	P
1TET	H	L	P	4P3C	H	L	M
1TJG	H	L	P	4Q0X	H	L	E
1TZG	H	L	P	4QXT	A	B	Q
1UWX	H	L	P	4RAV	A	B	E
1XGY	H	L	P	4TQE	H	L	A
2A6D	B	A	P	4TUJ	A	B	E
2B0S	H	L	P	4WHT	G	H	g
2BRR	H	L	P	4XGZ	A	B	a
2CK0	H	L	P	4XH2	A	B	a
2EH8	H	L	P	4XVJ	H	L	A
2G5B	B	A	I	4XXD	B	A	C
2GSI	B	A	W	4YDV	H	L	P
2H1P	H	L	P	4YHP	C	D	Q
2HFG	H	L	R	4YO0	A	B	E
2HH0	H	L	P	4YR6	A	B	C
2HKF	H	L	P	4Z0X	B	A	C
2HRP	H	L	P	4ZFO	A	B	F
2IGF	H	L	P	4ZTO	H	L	P
2IPU	G	K	P	5AUM	A	B	D
2J4W	H	L	D	5CSZ	A	B	E
2OQJ	B	A	C	5DD0	H	L	P
2OR9	H	L	P	5DLM	H	L	X
2OSL	A	B	Q	5DMG	E	F	X
2OTU	D	C	Q	5DRZ	B	A	Q
2QHR	H	L	P	5DS8	H	L	P
2QSC	H	L	P	5DSC	E	F	M
2V17	H	L	A	5E2V	H	L	P
2W65	A	B	E	5E8D	H	L	A

Continued on next page

Table D.2 – continued from previous page

PDB ID	H Chain	L Chain	Antigen Chain	PDB ID	H Chain	L Chain	Antigen Chain
2XZQ	H	L	P	5EA0	H	L	P
2Y5T	A	B	E | G | F	5EOC	H	L	P
2Y6S	D	C	P	5EOQ	H	L	A
2ZPK	H	L	P	5EOR	H	L	A
3BAE	H	L	A	5EPM	A	B	C
3BKY	H	L	P	5FGB	C	A	F
3CXD	H	L	P	5FGC	E	B	A
3E8U	H	L	P	5GIR	A	B	D
3EFD	H	L	K	5I8C	A	B	C
3EYF	B	A	E	5IFJ	A	B	C
3FFD	A	B	P	5IJK	A	C	X
3FN0	H	L	P	5IQ9	A	B	C
3G5V	B	A	C	5KN5	A	B	C
3G5Y	B	A	E	5MO3	H	L	A
3GGW	B	A	E	5MU0	A	B	Q
3GHE	H	L	P	5MY4	B	A	C
3GO1	H	L	P	5MYO	B	A	E
3H0T	B	A	C	5MYX	B	A	E
3HR5	B	A	S	5NPH	H	L	A
3IET	B	A	X	5T6P	B	A	F
3IFL	H	L	P	5TBD	A	B	C
3IFO	A	B	Q	5TKJ	A	B	C
3IFP	A	B	Q	5TKK	H	L	A
3IXT	A	B	C	5U3J	H	L	A
3LEX	A	B	C	5U3K	A	B	C
3LEY	H	L	P	5U3M	H	L	A
3MLR	H	L	P	5U3O	H	L	A
3MLW	H	L	P	5V6L	H	L	P
3MLX	H	L	P	5V6M	H	L	P
3MNW	B	A	P	5VXR	H	L	P
3O41	A	B	C	5VZY	H	L	A
3O6L	H	L	C	5W3P	H	L	P
3PP4	H	L	P	5WTT	A	B	C
3QG6	B	A	D	5XCS	A	B	C

D.3 Feature importance comparison between antibodies against proteins and peptides

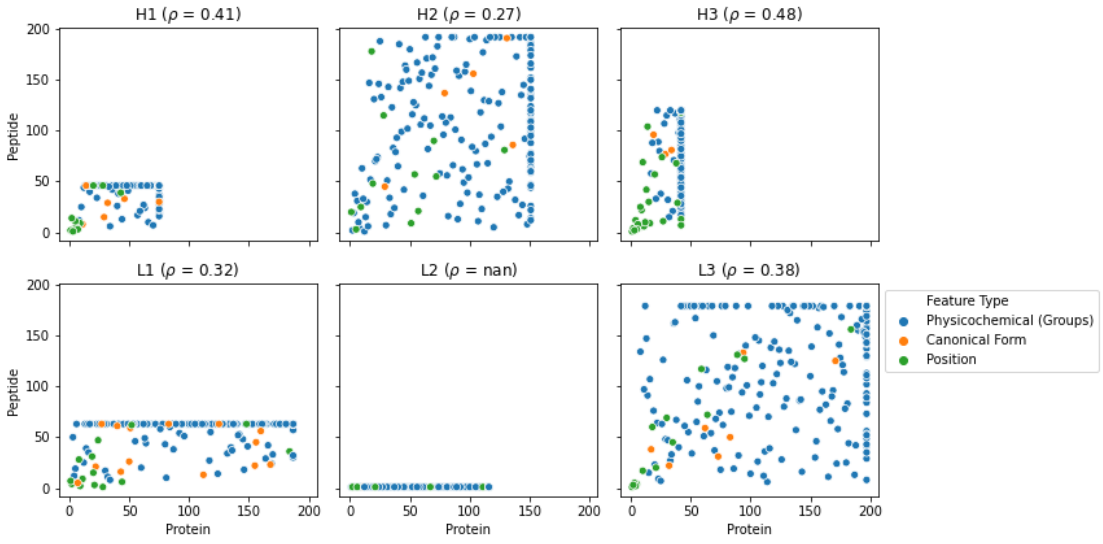


Figure D.1: Feature importance of the LASSO Triplet Groups model by CDR types, between protein-binding (Parapred set) and peptide-binding (peptide set) antibodies. See Figure 5.20 for the detailed caption.

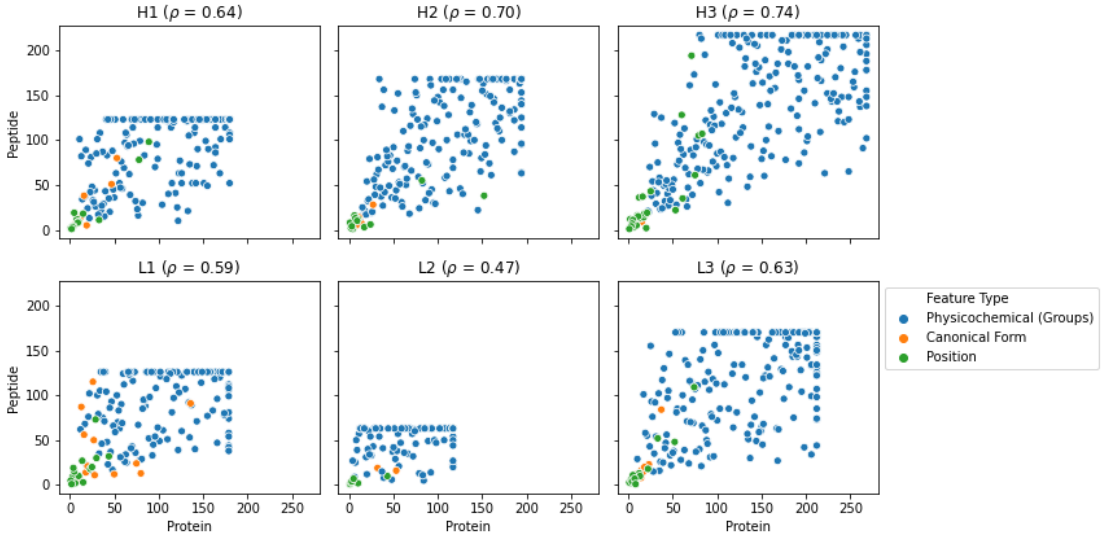


Figure D.2: Feature importance of the RF Triplet Groups model by CDR types, between protein-binding (Parapred set) and peptide-binding (peptide set) antibodies. See Figure 5.20 for the detailed caption.

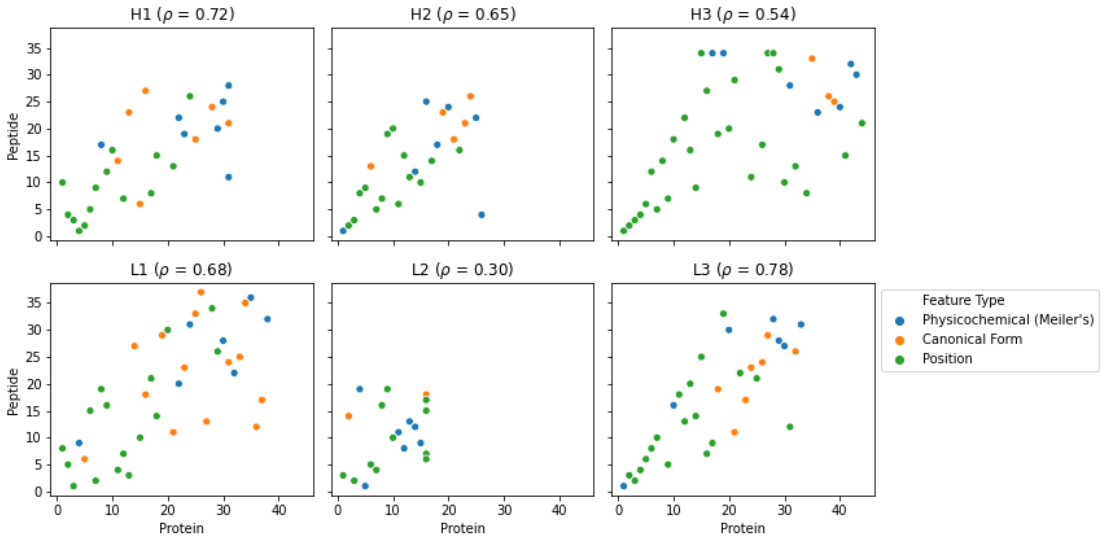


Figure D.3: Feature importance of the LASSO Single Meiler model by CDR types, between protein-binding (Parapred set) and peptide-binding (peptide set) antibodies. See Figure 5.20 for the detailed caption.

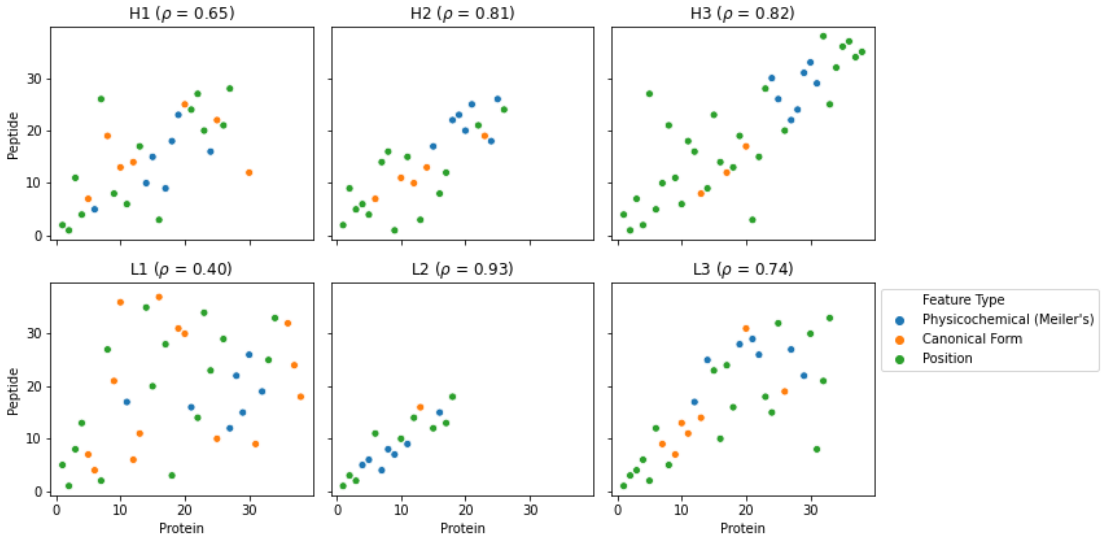


Figure D.4: Feature importance of the RF Single Meiler model by CDR types, between protein-binding (Parapred set) and peptide-binding (peptide set) antibodies. See Figure 5.20 for the detailed caption.

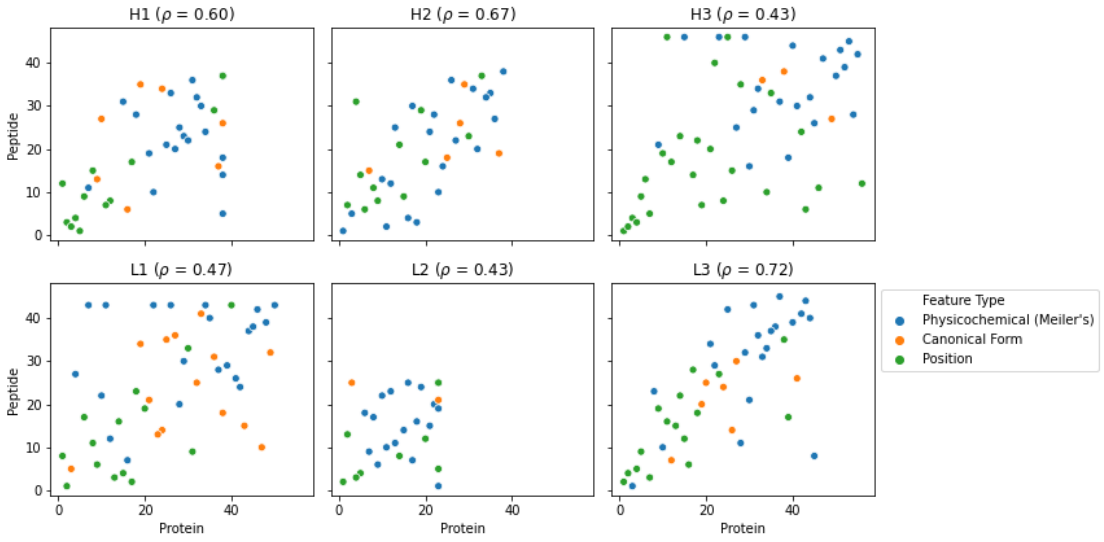


Figure D.5: Feature importance of the LASSO Triplet Meiler model by CDR types, between protein-binding (Parapred set) and peptide-binding (peptide set) antibodies. See Figure 5.20 for the detailed caption.

References

W. M. Abbott, M. M. Damschroder, and D. C. Lowe. Current approaches to fine mapping of antigen–antibody interactions. *Immunology*, 142(4):526–535, 2014.

Y. N. Abdiche, A. Y. Yeung, I. Ni, D. Stone, A. Miles, W. Morishige, A. Rossi, and P. Strop. Antibodies targeting closely adjacent or minimally overlapping epitopes can displace one another. *PLoS One*, 12(1):e0169535, 2017.

R. M. Adams, T. Mora, A. M. Walczak, and J. B. Kinney. Measuring the sequence-affinity landscape of antibodies with massively parallel titration curves. *eLife*, 5:e23156, 2016.

B. Al-Lazikani, A. M. Lesk, and C. Chothia. Standard conformations for the canonical structures of immunoglobulins. *Journal of Molecular Biology*, 273(4):927–948, 1997.

B. Al-Lazikani, A. M. Lesk, and C. Chothia. Canonical structures for the hypervariable regions of T cell $\alpha\beta$ receptors. *Journal of Molecular Biology*, 295(4):979–995, 2000.

B. Alberts, A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter. The Adaptive Immune System. In *Molecular Biology of the Cell*, chapter 24, pages 1363–1422. Garland Science, New York, 4th edition, 2002.

L. P. Albou, B. Schwarz, O. Poch, J. M. Wurtz, and D. Moras. Defining and characterizing protein surface using alpha shapes. *Proteins: Structure, Function and Bioinformatics*, 76(1):1–12, 2009.

T. J. Allison, C. C. Winter, J.-J. Fournié, M. Bonneville, and D. N. Garboczi. Structure of a human $\gamma\delta$ T-cell antigen receptor. *Nature*, 411:820–824, 2001.

J. C. Almagro, A. Teplyakov, J. Luo, R. W. Sweet, S. Kodangattil, F. Hernandez-Guzman, and G. L. Gilliland. Second Antibody Modeling Assessment (AMA-II). *Proteins: Structure, Function and Bioinformatics*, 82(8):1553–1562, 2014.

A. Altmann, L. Toloşi, O. Sander, and T. Lengauer. Permutation importance: a corrected feature importance measure. *Bioinformatics*, 26(10):1340–1347, 2010.

F. Ambrosetti, T. H. Olsen, P. P. Olimpieri, B. Jiménez-García, E. Milanetti, P. Marcatilli, and A. M. J. J. Bonvin. proABC-2: PRediction of AntiBody contacts v2 and its application to information-driven docking. *Bioinformatics*, page btaa644, 7 2020.

M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points to Identify the Clustering Structure. *ACM SIGMOD Record*, 28(2):49–60, 1999.

M. S. Armstrong, G. M. Morris, P. W. Finn, R. Sharma, L. Moretti, R. I. Cooper, and W. G. Richards. ElectroShape: Fast molecular similarity calculations incorporating shape, chirality and electrostatics. *Journal of Computer-Aided Molecular Design*, 24 (9):789–801, 2010.

M. Attaf, E. Huseby, and A. K. Sewell. $\alpha\beta$ T cell receptors as predictors of health and disease. *Cellular & Molecular Immunology*, 12:391–399, 2015.

Y. Avnir, A. S. Tallarico, Q. Zhu, A. S. Bennett, G. Connelly, J. Sheehan, J. Sui, A. Fahmy, C. yu Huang, G. Cadwell, L. A. Bankston, A. T. McGuire, L. Stamatatos, G. Wagner, R. C. Liddington, and W. A. Marasco. Molecular Signatures of Hemagglutinin Stem-Directed Heterosubtypic Human Neutralizing Antibodies against Influenza A Viruses. *PLoS Pathogens*, 10(5):e1004103, 2014.

P. J. Ballester and W. G. Richards. Ultrafast shape recognition for similarity search in molecular databases. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 463(2081):1307–1321, 2007.

Y. Barrios, P. Jirholt, and M. Ohlin. Length of the antibody heavy chain complementarity determining region 3 as a specificity-determining factor. *Journal of Molecular Recognition*, 17(4):332–338, 2004.

G. Bentley, G. Boulot, and R. Mariuzza. The structure of the antigen-binding site of immunoglobulins and T-cell receptors. *Research in Immunology*, 146(4):277–290, 1995.

H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28: 235–242, 2000.

M. Biasini, S. Bienert, A. Waterhouse, K. Arnold, G. Studer, T. Schmidt, F. Kiefer, T. G. Cassarino, M. Bertoni, L. Bordoli, and T. Schwede. SWISS-MODEL: Modelling protein tertiary and quaternary structure using evolutionary information. *Nucleic Acids Research*, 42(W1):W252–W258, 2014.

S. J. Blevins, B. G. Pierce, N. K. Singh, T. P. Riley, Y. Wang, T. T. Spear, M. I. Nishimura, Z. Weng, and B. M. Baker. How structural adaptability exists alongside HLA-A2 bias in the human $\alpha\beta$ TCR repertoire. *Proceedings of the National Academy of Sciences*, 113(9):E1276–E1285, 2016.

J. Bostrom, S.-F. Yu, D. Kan, B. A. Appleton, C. V. Lee, K. Billeci, W. Man, F. Peale, S. Ross, C. Wiesmann, and G. Fuh. Variants of the antibody herceptin that interact with HER2 and VEGF at the antigen binding site. *Science*, 323(5921):1610–1614, 2009.

D. A. Cannon, L. Shan, Q. Du, L. Shirinian, K. W. Rickert, K. L. Rosenthal, M. Korade III, L. E. van Vlerken-Ysla, A. Buchanan, T. J. Vaughan, M. M. Damschroder, and B. Popovic. Experimentally guided computational antibody affinity maturation with *de novo* docking modelling and rational design. *PLoS Computational Biology*, 15(5):e1006980, 2019.

M. M. Castellanos, J. A. Snyder, M. Lee, S. Chakravarthy, N. J. Clark, A. McAuley, and J. E. Curtis. Characterization of monoclonal antibody–protein antigen complexes using small-angle scattering and molecular modeling. *Antibodies*, 6(4):25, 2017.

P. Chames, M. Van Regenmortel, E. Weiss, and D. Baty. Therapeutic antibodies: Successes, limitations and hopes for the future. *British Journal of Pharmacology*, 157 (2):220–233, 2009.

G. M. Cherf and J. R. Cochran. Applications of Yeast Surface Display for Protein Engineering. In B. Liu, editor, *Yeast Surface Display: Methods, Protocols, and Applications*, pages 155–175. Springer New York, New York, NY, 2015.

M. . Chiu, D. R. Goulet, A. Teplyakov, and G. L. Gilliland. Antibody structure and function: The basis for engineering therapeutics. *Antibodies*, 8(4):55, 2019.

Y. Choi and C. M. Deane. FREAD revisited: Accurate loop structure prediction using a database search algorithm. *Proteins*, 78(6):1431–1440, 2010.

C. Chothia and A. M. Lesk. Canonical structures for the hypervariable regions of immunoglobulins. *Journal of Molecular Biology*, 196(4):901–917, 1987.

P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423, 2009.

D. K. Cole, K. M. Miles, F. Madura, C. J. Holland, A. J. A. Schauenburg, A. J. Godkin, A. M. Bulek, A. Fuller, H. J. E. Akpovwa, P. G. Pymm, N. Liddy, M. Sami, Y. Li, P. J. Rizkallah, B. K. Jakobsen, and A. K. Sewell. T-cell Receptor (TCR)-Peptide Specificity Overrides Affinity-enhancing TCR-Major Histocompatibility Complex Interactions. *Journal of Biological Chemistry*, 289(2):628–638, 2014.

G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004.

J. E. Crooks, C. T. Boughter, L. R. Scott, and E. J. Adams. The hypervariable loops of free TCRs sample multiple distinct metastable conformations in solution. *Frontiers in Molecular Biosciences*, 5:95, 2018.

S. Daberdaku and C. Ferrari. Antibody interface prediction with 3D Zernike descriptors and SVM. *Bioinformatics*, 35(11):1870–1876, 2019.

A. Deac, P. Veličković, and P. Sormanni. Attentive cross-modal paratope prediction. *Journal of Computational Biology*, 26(6):536–545, 2019.

C. M. Deane and T. L. Blundell. CODA: a combined algorithm for predicting the structurally variable regions of protein models. *Protein Science*, 10(3):599–612, 2001.

B. J. DeKosky, O. I. Lungu, D. Park, E. L. Johnson, W. Charab, C. Chrysostomou, D. Kuroda, A. D. Ellington, G. C. Ippolito, J. J. Gray, and G. Georgiou. Large-scale sequence and structural comparisons of human naive and antigen-experienced antibody repertoires. *Proceedings of the National Academy of Sciences*, 113(19):E2636–E2645, 2016.

L. Di Rienzo, E. Milanetti, R. Lepore, P. P. Olimpieri, and A. Tramontano. Superposition-free comparison and clustering of antibody binding sites: implications for the prediction of the nature of their antigen. *Scientific Reports*, 7(1):1–10, 2017.

L. Dubrovsky, T. Dao, R. S. Gejman, E. J. Brea, A. Y. Chang, C. Y. Oh, E. Casey, D. Pankov, and S. D. A. T cell receptor mimic antibodies for cancer therapy. *OncoImmunology*, 5(1):e1049803, 2016.

J. Dunbar and C. M. Deane. ANARCI: Antigen receptor numbering and receptor classification. *Bioinformatics*, 32(2):298–300, 2016.

J. Dunbar, B. Knapp, A. Fuchs, J. Shi, and C. M. Deane. Examining Variable Domain Orientations in Antigen Receptors Gives Insight into TCR-Like Antibody Design. *PLoS Computational Biology*, 10(9):1–10, 2014a.

J. Dunbar, K. Krawczyk, J. Leem, T. Baker, A. Fuchs, G. Georges, J. Shi, and C. M. Deane. SAbDab: the structural antibody database. *Nucleic Acids Research*, 42(D1): D1140–D1146, 2014b.

J.-P. Ebejer, P. W. Finn, W. K. Wong, C. M. Deane, and G. M. Morris. Ligity: A Non-Superpositional, Knowledge-Based Approach to Virtual Screening. *Journal of Chemical Information and Modeling*, 59(6):2600–2616, 2019.

M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A Density-based Algorithm for Discovering Clusters a Density-based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, pages 226–231. AAAI Press, 1996.

M. L. Fernández-Quintero, J. Kraml, G. Georges, and K. R. Liedl. CDR-H3 loop ensemble in solution – conformational selection upon antibody binding. *mAbs*, 11(6): 1077–1088, 2019a.

M. L. Fernández-Quintero, J. R. Loeffler, J. Kraml, U. Kahler, A. S. Kamenik, and K. R. Liedl. Characterizing the Diversity of the CDR-H3 Loop Conformational Ensembles in Relationship to Antibody Binding Properties. *Frontiers in Immunology*, 9:3065, 2019b.

M. L. Fernández-Quintero, J. R. Loeffler, L. M. Bacher, F. Waibl, C. A. Seidler, and K. R. Liedl. Local and Global Rigidification Upon Antibody Affinity Maturation. *Frontiers in Molecular Biosciences*, 7:182, 2020a.

M. L. Fernández-Quintero, N. D. Pomarici, J. R. Loeffler, C. A. Seidler, and K. R. Liedl. T-Cell Receptor CDR3 Loop Conformations in Solution Shift the Relative Vα-Vβ Domain Distributions. *Frontiers in Immunology*, 11:1440, 2020b.

J. A. Finn, J. Koehler Leman, J. R. Willis, A. Cisneros III, J. E. Crowe Jr, and J. Meiler. Improving loop modeling of the antibody complementarity-determining region 3 using knowledge-based restraints. *PLoS One*, 11(5):e0154811, 2016.

J. R. Francica, Z. Sheng, Z. Zhang, Y. Nishimura, M. Shingai, A. Ramesh, B. F. Keele, S. D. Schmidt, B. J. Flynn, S. Darko, R. M. Lynch, T. Yamamoto, R. Matus-Nicodemos, D. Wolinsky, M. Nason, N. M. Valiante, P. Malyala, E. De Gregorio, S. W. Barnett, M. Singh, D. T. O'Hagan, R. A. Koup, J. R. Mascola, M. A. Martin, T. B. Kepler, D. C. Douek, L. Shapiro, R. A. Seder, B. Barnabas, R. Blakesley, G. Bouffard, S. Brooks, H. Coleman, M. Dekhtyar, M. Gregory, X. Guan, J. Gupta, J. Han, S. L. Ho, R. Legaspi, Q. Maduro, C. Masiello, B. Maskeri, J. McDowell, C. Montemayor, J. Mullikin, M. Park, N. Riebow, K. Schandler, B. Schmidt, C. Sison,

M. Stantripop, J. Thomas, P. Thomas, M. Vemulapalli, and A. Young. Analysis of immunoglobulin transcripts and hypermutation following SHIV$_{AD8}$ infection and protein-plus-adjuvant immunization. *Nature Communications*, 6(1):6565, 2015.

P. Gainza, F. Sverrisson, F. Monti, E. Rodolà, D. Boscaini, M. M. Bronstein, and B. E. Correia. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, 17(2):184–192, 2020.

J. D. Galson, J. Trück, A. Fowler, M. Münz, V. Cerundolo, A. J. Pollard, G. Lunter, and D. F. Kelly. In-depth assessment of within-individual and inter-individual variation in the B cell receptor repertoire. *Frontiers in Immunology*, 6:531, 2015.

M. Gao and J. Skolnick. APoc: large-scale identification of similar protein pockets. *Bioinformatics*, 29(5):597–604, 2013.

K. C. Garcia, M. Degano, L. R. Pease, M. Huang, P. A. Peterson, L. Teyton, and I. A. Wilson. Structural basis of plasticity in T cell receptor recognition of a self peptide-MHC antigen. *Science*, 279(5354):1166–1172, 1998.

A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. Paula Magariños, J. P. Overington, G. Papadatos, I. Smit, and A. R. Leach. The ChEMBL database in 2017. *Nucleic Acids Research*, 45(D1):D945–D954, 2017.

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

C. Göbl, T. Madl, B. Simon, and M. Sattler. NMR approaches for structural analysis of multidomain proteins and complexes in solution. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 80:26–63, 2014.

R. Gowthaman and B. G. Pierce. TCRmodel: high resolution modeling of T cell receptors from sequence. *Nucleic Acids Research*, 46(W1):W396–W401, 2018.

A. J. Greaney, T. N. Starr, P. Gilchuk, S. J. Zost, E. Binshtein, A. N. Loes, S. K. Hilton, J. Huddleston, R. Eguia, K. H. D. Crawford, A. S. Dingens, R. S. Nargi, R. E. Sutton, N. Suryadevara, P. W. Rothlauf, Z. Liu, S. P. Whelan, R. H. Carnahan, J. E. Crowe Jr., and J. D. Bloom. Complete mapping of mutations to the SARS-CoV-2 spike receptor-binding domain that escape antibody recognition. *BioRxiv*, 2020.

S. Henikoff and J. G. Henikoff. Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89(22):10915–10919, 1992.

T. Hoffmann, A. Marion, and I. Antes. DynaDom: structure-based prediction of T cell receptor inter-domain and T cell receptor-peptide-MHC (class I) association angles. *BMC Structural Biology*, 17(1):2, 2018.

C. J. Holland, B. J. MacLachlan, V. Bianchi, S. J. Hesketh, R. Morgan, O. Vickery, A. M. Bulek, A. Fuller, A. Godkin, A. K. Sewell, P. J. Rizkallah, S. Wells, and D. K. Cole. In Silico and structural analyses demonstrate that intrinsic protein motions guide T cell

receptor complementarity determining region loop flexibility. *Frontiers in Immunology*, 9:674, 2018.

L. Holm and C. Sander. Dali: a network tool for protein structure comparison. *Trends in Biochemical Sciences*, 20(11):478–480, 1995.

Y.-C. Hsiao, Y.-J. J. Chen, L. D. Goldstein, J. Wu, Z. Lin, K. Schneider, S. Chaudhuri, A. Antony, K. B. Pahuja, Z. Modrusan, D. Seshasayee, S. Seshagiri, and I. Hötzel. Restricted epitope specificity determined by variable region germline segment pairing in rodent antibody repertoires. *mAbs*, 12(1):1722541, 2020.

R. Y.-C. Huang, M. Kuhne, S. Deshpande, V. Rangan, M. Srinivasan, Y. Wang, and G. Chen. Mapping binding epitopes of monoclonal antibodies targeting major histocompatibility complex class I chain-related A (MICA) with hydrogen/deuterium exchange and electron-transfer dissociation mass spectrometry. *Analytical and Bioanalytical Chemistry*, 412(7):1693–1700, 2020.

C. A. Janeway, P. Travers, M. Walport, and M. J. Shlomchik. *Immunobiology: the immune system in health and disease*. Garland Science, 5th edition, 2001.

J. R. Jeliazkov, A. Sljoka, D. Kuroda, N. Tsuchimura, N. Katoh, K. Tsumoto, and J. J. Gray. Repertoire analysis of antibody CDR-H3 loops suggests affinity maturation does not typically result in rigidification. *Frontiers in Immunology*, 9:413, 2018.

K. K. Jensen, V. Rantos, E. C. Jappe, T. H. Olsen, M. C. Jespersen, V. Jurtz, L. E. Jessen, E. Lanzarotti, S. Mahajan, B. Peters, M. Nielsen, and P. Marcatili. TCRpMHCmodels: Structural modelling of TCR-pMHC class I complexes. *Scientific Reports*, 9(1):1–12, 2019.

H. Jubb, A. Higueruelo, B. Ochoa-Montaño, W. R. Pitt, D. Ascher, and B. T. Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures. *Journal of Molecular Biology*, 429(3):365–371, 2017.

E. A. Kabat and T. T. Wu. Identical V region amino acid sequences and segments of sequences in antibodies of different specificities. Relative contributions of VH and VL genes, minigenes, and complementarity-determining regions to binding of antibody-combining sites. *Journal of Immunology*, 147(5):1709–1719, 1991.

S. Kawashima, P. Pokarowski, M. Pokarowska, A. Kolinski, T. Katayama, and M. Kanehisa. AAindex: amino acid index database, progress report 2008. *Nucleic Acids Research*, 36(Suppl_1):D202–D205, 2007.

M. S. Klausen, M. V. Anderson, M. C. Jespersen, M. Nielsen, and P. Marcatili. LYRA, a webserver for lymphocyte receptor structural modeling. *Nucleic Acids Research*, 43 (W1):W349–W355, 2015.

P. Koenig, C. V. Lee, S. Sanowar, P. Wu, J. Stinson, S. F. Harris, and G. Fuh. Deep sequencing-guided design of a high affinity dual specificity antibody to target two angiogenic factors in neovascular age-related macular degeneration. *Journal of Biological Chemistry*, 290(36):21773–21786, 2015.

P. Koenig, C. V. Lee, B. T. Walters, V. Janakiraman, J. Stinson, T. W. Patapoff, and G. Fuh. Mutational landscape of antibody variable domains reveals a switch modulating the interdomain conformational dynamics and antigen binding. *Proceedings of the National Academy of Sciences*, 114(4):E486–E495, 2017.

A. Kovaltsuk, K. Krawczyk, J. D. Galson, D. F. Kelly, C. M. Deane, and J. Trück. How B-cell receptor repertoire sequencing can be enriched with structural antibody data. *Frontiers in Immunology*, 8:1753, 2017.

A. Kovaltsuk, J. Leem, S. Kelm, J. Snowden, C. M. Deane, and K. Krawczyk. Observed antibody space: A resource for data mining next-generation sequencing of antibody repertoires. *The Journal of Immunology*, 201(8):2502–2509, 2018.

A. Kovaltsuk, M. I. J. Raybould, W. K. Wong, C. Marks, S. Kelm, J. Snowden, J. Trück, and C. M. Deane. Structural diversity of B-cell receptor repertoires along the B-cell differentiation axis in humans and mice. *PLoS Computational Biology*, 16(2):e1007636, 2020.

J. C. Krause, T. Tsibane, T. M. Tumpey, C. J. Huffman, B. S. Briney, S. A. Smith, C. F. Basler, and J. E. Crowe. Epitope-specific human influenza antibody repertoires diversify by B cell intraclonal sequence divergence and interclonal convergence. *The Journal of Immunology*, 187(7):3704–3711, 2011.

K. Krawczyk, T. Baker, J. Shi, and C. M. Deane. Antibody i-Patch prediction of the antibody binding site improves rigid local antibody-antigen docking. *Protein Engineering Design and Selection*, 26(10):621–629, 2013.

K. Krawczyk, X. Liu, T. Baker, J. Shi, and C. M. Deane. Improving B-cell epitope prediction and its application to global antibody-antigen docking. *Bioinformatics*, 30 (16):2288–2294, 2014.

K. Krawczyk, S. Kelm, A. Kovaltsuk, J. D. Galson, D. Kelly, J. Trück, C. Regep, J. Leem, W. K. Wong, J. Nowak, J. Snowden, M. Wright, L. Starkie, A. Scott-Tucker, J. Shi, and C. M. Deane. Structurally mapping antibody repertoires. *Frontiers in Immunology*, 9:1698, 2018.

J. V. Kringelum, M. Nielsen, S. B. Padkjær, and O. Lund. Structural analysis of B-cell epitopes in antibody: protein complexes. *Molecular Immunology*, 53(1-2):24–34, 2013.

L. A. Kuhn, M. A. Siani, M. E. Pique, C. L. Fisher, E. D. Getzoff, and J. A. Tainer. The interdependence of protein surface topography and bound water molecules revealed by surface accessibility and fractal density measures. *Journal of Molecular Biology*, 228 (1):13–22, 1992.

V. Kunik, S. Ashkenazi, and Y. Ofran. Paratome: An online tool for systematic identification of antigen-binding regions in antibodies based on sequence or structure. *Nucleic Acids Research*, 40(W1):W521–W524, 2012a.

V. Kunik, B. Peters, and Y. Ofran. Structural consensus among antibodies defines the antigen binding site. *PLoS Computational Biology*, 8(2):e1002388, 2012b.

D. Kuroda, H. Shirai, M. Kobori, and H. Nakamura. Structural classification of CDR-H3 revisited: a lesson in antibody modeling. *Proteins: Structure, Function, and Bioinformatics*, 73(3):608–620, 2008.

D. Kuroda, H. Shirai, M. Kobori, and H. Nakamura. Systematic classification of CDR–L3 in antibodies: Implications of the light chain subtypes and the VL–VH interface. *Proteins*, 75(1):139–146, 2009.

J.-W. Kwak and C.-S. Yoon. A convenient method for epitope competition analysis of two monoclonal antibodies for their antigen binding. *Journal of Immunological Methods*, 191(1):49–54, 1996.

G. D. Lapidoth, D. Baran, G. M. Pszolla, C. Norn, A. Alon, M. D. Tyka, and S. J. Fleishman. AbDesign: An algorithm for combinatorial backbone design guided by natural conformations and sequences. *Proteins*, 83(8):1385–1406, 2015.

H. Y. Lee, G. Schaefer, I. Lesaca, C. V. Lee, P. Y. Wong, and G. Jiang. "Two-in-one" approach for bioassay selection for dual specificity antibodies. *Journal of Immunological Methods*, 448:74–79, 2017.

M. Lee, P. Lloyd, X. Zhang, J. M. Schallhorn, K. Sugimoto, A. G. Leach, G. Sapiro, and K. N. Houk. Shapes of Antibody Binding Sites: Qualitative and Quantitative Analyses Based on a Geomorphic Classification Scheme. *The Journal of Organic Chemistry*, 71 (14):5082–5092, 2006.

J. Leem, J. Dunbar, G. Georges, J. Shi, and C. M. Deane. ABodyBuilder: Automated antibody structure prediction with data–driven accuracy estimation. *mAbs*, 8(7): 1259–1268, 2016.

J. Leem, S. H. P. de Oliveira, K. Krawczyk, and C. M. Deane. STCRDab: the structural T-cell receptor database. *Nucleic Acids Research*, 46(D1):D406–D412, 2018a.

J. Leem, G. Georges, J. Shi, and C. M. Deane. Antibody side chain conformations are position-dependent. *Proteins: Structure, Function, and Bioinformatics*, 86(4):383–392, 2018b.

M.-P. Lefranc, V. Giudicelli, P. Duroux, J. Jabado-Michaloud, G. Folch, S. Aouinti, E. Carillon, H. Duvergey, A. Houles, T. Paysan-Lafosse, S. Hadi-Saljoqi, S. Sasorith, G. Lefranc, and S. Kossida. Imgt®, the international immunogenetics information system® 25 years on. *Nucleic Acids Research*, 43(D1):D413–D422, 2015.

M. F. Lensink, N. Nadzirin, S. Velankar, and S. J. Wodak. Modeling protein-protein, protein-peptide and protein-oligosaccharide complexes: CAPRI 7th edition. *Proteins: Structure, Function, and Bioinformatics*, 87:1200–1221, 2019.

S. Li, J. Wilamowski, S. Teraguchi, F. J. van Eerden, J. Rozewicki, A. Davila, Z. Xu, K. Katoh, and D. M. Standley. Structural Modeling of Lymphocyte Receptors and Their Antigens. In *In Vitro Differentiation of T-Cells*, pages 207–229. Springer, 2019.

Y. Li, S. O'Dell, R. Wilson, X. Wu, S. D. Schmidt, C.-M. Hogerkorp, M. K. Louder, N. S. Longo, C. Poulsen, J. Guenaga, B. K. Chakrabarti, N. Doria-Rose, M. Roederer, M. Connors, J. R. Mascola, and R. T. Wyatt. HIV-1 neutralizing antibodies display

dual recognition of the primary and coreceptor binding sites and preferential binding to fully cleaved envelope glycoproteins. *Journal of Virology*, 86(20):11231–11241, 2012.

J. Liang, H. Edelsbrunner, and C. Woodward. Anatomy of protein pockets and cavities: Measurement of binding site geometry and implications for ligand design. *Protein Science*, 7(9):1884–1897, 1998.

E. Liberis, P. Veličković, P. Sormanni, M. Vendruscolo, and P. Liò. Parapred: Antibody Paratope Prediction using Convolutional and Recurrent Neural Networks. *Bioinformatics*, 34(17):2944–2950, 2018.

C. C. Lim, Y. S. Choong, and T. S. Lim. Cognizance of molecular methods for the generation of mutagenic phage display antibody libraries for affinity maturation. *International Journal of Molecular Sciences*, 20(8):1861, 2019.

R. M. MacCallum, A. C. Martin, and J. M. Thornton. Antibody-antigen interactions: Contact analysis and binding site topography. *Journal of Molecular Biology*, 262(5): 732–745, 1996.

A. R. Maceiras, S. Almeida, E. Mariotti-Ferrandiz, W. Chaara, A. Six, S. Hori, D. Klatzmann, J. Faro, and L. Graca. T follicular helper and T follicular regulatory cells have different TCR specificity. *Nature Communications*, 8(15067), 2017.

T. Magoč and S. L. Salzberg. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics*, 27(21):2957–2963, 2011.

J. K. Maier and P. Labute. Assessment of fully automated antibody homology modeling protocols in molecular operating environment. *Proteins*, 82(8):1599–1610, 2014.

A.-C. Malmborg and C. A. K. Borrebaeck. BIAcore as a tool in antibody engineering. *Journal of Immunological methods*, 183(1):7–13, 1995.

C. Marks and C. M. Deane. Antibody H3 Structure Prediction. *Computational and Structural Biotechnology Journal*, pages 222–231, 2017.

C. Marks and C. M. Deane. How repertoire data is changing antibody science. *Journal of Biological Chemistry*, 295:9823–9837, 2020.

C. Marks, J. Nowak, S. Klostermann, G. Georges, J. Dunbar, J. Shi, S. Kelm, and C. M. Deane. Sphinx: Merging knowledge-based and ab initio approaches to improve protein loop prediction. *Bioinformatics*, 33(9):1346–1353, 2017.

A. C. R. Martin and J. M. Thornton. Structural families in loops of homologous proteins: automatic classification, modelling and application to antibodies. *Journal of Molecular Biology*, 263(5):800–815, 1996.

D. M. Mason, S. Friedensohn, C. R. Weber, C. Jordi, B. Wagner, S. Meng, P. Gainza, B. Correia, and S. T. Reddy. Deep learning enables therapeutic antibody optimization in mammalian cells by deciphering high-dimensional protein sequence space. *BioRxiv*, 2019.

J. Meiler, M. Müller, A. Zeidler, and F. Schmäschke. Generation and evaluation of dimension-reduced amino acid parameter representations by artificial neural networks. *Molecular Modeling Annual*, 7(9):360–369, 2001.

C. Mirabello and B. Wallner. Topology independent structural matching discovers novel templates for protein interfaces. *Bioinformatics*, 34(17):i787–i794, 2018.

A. K. Mishra and R. A. Mariuzza. Insights into the structural basis of antibody affinity maturation from next-generation sequencing. *Frontiers in Immunology*, 9:117, 2018.

S. Mohan, N. Sinha, and S. Smith-Gill. Modeling the binding sites of anti-hen egg white lysozyme antibodies HyHEL-8 and HyHEL-26: an insight into the molecular basis of antibody cross-reactivity and specificity. *Biophysical Journal*, 85(5):3221–36, 2003.

V. Morea, A. Tramontano, M. Rustici, C. Chothia, and A. M. Lesk. Conformations of the third hypervariable region in the VH domain of immunoglobulins. *Journal of Molecular Biology*, 275(2):269–294, 1998.

G. P. Morris and P. M. Allen. How the TCR balances sensitivity and specificity for the recognition of self and pathogens. *Nature Immunology*, 13(2):121, 2012.

T. A. Najar, S. Khare, R. Pandey, S. K. Gupta, and R. Varadarajan. Mapping Protein Binding Sites and Conformational Epitopes Using Cysteine Labeling and Yeast Surface Display. *Structure*, 25(3):395–406, 2017.

M. N. Nguyen, K. P. Tan, and M. S. Madhusudhan. CLICK - Topology-independent comparison of biomolecular 3D structures. *Nucleic Acids Research*, 39(Suppl_2): W24–W28, 2011.

M. N. Nguyen, M. R. Pradhan, C. Verma, and P. Zhong. The interfacial character of antibody paratopes: analysis of antibody–antigen structures. *Bioinformatics*, 33(19): 2971–2976, 2017.

B. North, A. Lehmann, and R. L. Dunbrack Jr. A new clustering of antibody cdr loop conformations. *Journal of Molecular Biology*, 406(2):228–256, 2011.

J. Nowak, T. Baker, G. Georges, S. Kelm, S. Klostermann, J. Shi, S. Sridharan, and C. M. Deane. Length-independent structural similarities enrich the antibody CDR canonical class model. *mAbs*, 8(4):751–760, 2016.

G. Núñez-Vivanco, A. Valdés-Jiménez, F. Besoaín, and M. Reyes-Parada. Geomfinder: A multi-feature identifier of similar three-dimensional protein patterns: A ligand-independent approach. *Journal of Cheminformatics*, 8(1):19, 2016.

P. P. Olimpieri, A. Chailyan, A. Tramontano, and P. Marcatili. Prediction of site-specific interactions in antibody-antigen complexes: The proABC method and server. *Bioinformatics*, 29(18):2285–2291, 2013.

M. A. Olson, P. M. Legler, D. Zabetakis, K. B. Turner, G. P. Anderson, and E. R. Goldman. Sequence Tolerance of a Single-Domain Antibody with a High Thermal Stability: Comparison of Computational and Experimental Fitness Profiles. *ACS Omega*, 4(6):10444–10454, 2019.

T. W. Pai, H. W. Wang, Y. C. Lin, and H. T. Chang. Prediction of B-cell linear epitopes with a combination of support vector machine classification and amino acid propensity identification. *BioMed Research International*, 2011:432830, 2011.

J. Pan, S. Zhang, A. Chou, D. B. Hardie, and C. H. Borchers. Fast comparative structural characterization of intact therapeutic antibodies using hydrogen–deuterium exchange and electron transfer dissociation. *Analytical Chemistry*, 87(12):5884–5890, 2015.

R. J. Pantazes and C. D. Maranas. OptCDR: A general computational method for the design of antibody complementarity determining regions for targeted epitope binding. *Protein Engineering, Design and Selection*, 23(11):849–858, 2010.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and É. Duchesnay. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

B. G. Pierce and Z. Weng. A flexible docking approach for prediction of T cell receptor–peptide–MHC complexes. *Protein Science*, 22(1):35–46, 2013.

S. Pittala and C. Bailey-Kellogg. Learning context-aware structural representations to predict antigen and antibody binding interfaces. *Bioinformatics*, 36(13):3996–4003, 2020.

C. Pommié, S. Levadoux, R. Sabatier, G. Lefranc, and M.-P. Lefranc. IMGT standardized criteria for statistical analysis of immunoglobulin V-REGION amino acid properties. *Journal of Molecular Recognition*, 17(1):17–32, 2004.

J. Pons, J. R. Stratton, and J. F. Kirsch. How do two unrelated antibodies, HyHEL-10 and F9. 13.7, recognize the same epitope of hen egg-white lysozyme? *Protein Science*, 11(10):2308–2315, 2002.

C. Puchades, B. Kűkrer, O. Diefenbach, E. Sneekes-Vriese, J. Juraszek, W. Koudstaal, and A. Apetri. Epitope mapping of diverse influenza Hemagglutinin drug candidates using HDX-MS. *Scientific Reports*, 9(1):4735, 2019.

T. Qiu, H. Xiao, Q. Zhang, J. Qiu, Y. Yang, D. Wu, Z. Cao, and R. Zhu. Proteochemometric modeling of the antigen-antibody interaction: New fingerprints for antigen, antibody and epitope-paratope interaction. *PLoS ONE*, 10(4):e0122416, 2015.

T. Ramaraj, T. Angel, E. A. Dratz, A. J. Jesaitis, and B. Mumey. Antigen–antibody interface properties: Composition, residue interactions, and features of 53 non-redundant structures. *Biochimica et Biophysica Acta (BBA)–Proteins and Proteomics*, 1824(3):520–532, 2012.

M. I. J. Raybould, C. Marks, K. Krawczyk, B. Taddese, J. Nowak, A. P. Lewis, A. Bujotzek, J. Shi, and C. M. Deane. Five computational developability guidelines for therapeutic antibody profiling. *Proceedings of the National Academy of Sciences*, 116 (10):4025–4030, 2019a.

M. I. J. Raybould, W. K. Wong, and C. M. Deane. Antibody–antigen complex modelling in the era of immunoglobulin repertoire sequencing. *Molecular Systems Design & Engineering*, 4(4):679–688, 2019b.

M. I. J. Raybould, C. Marks, A. Kovaltsuk, A. P. Lewis, J. Shi, and C. M. Deane. Evidence of Antibody Repertoire Functional Convergence through Public Baseline and Shared Response Structures. *BioRxiv*, 2020.

E. Richardson, J. D. Galson, P. Kellam, D. F. Kelly, S. E. Smith, A. Palser, S. Watson, and C. M. Deane. A computational method for immune repertoire mining that identifies novel binders from different clonotypes, demonstrated by identifying anti-Pertussis toxoid antibodies. *BioRxiv*, 2020.

E. P. Rock, P. R. Sibbald, M. M. Davis, and Y. H. Chien. CDR3 length in antigen-specific immune receptors. *Journal of Experimental Medicine*, 179(1):323–328, 1994.

J. Rossjohn, S. Gras, J. J. Miles, S. J. Turner, D. I. Godfrey, and J. McCluskey. T cell antigen receptor recognition of antigen-presenting molecules. *Annual Review of Immunology*, 33:169–200, 2015.

M. G. Rudolph, R. L. Stanfield, and I. A. Wilson. How TCRs bind MHCs, peptides, and coreceptors. *Annual Review of Immunology*, 24(1):419–466, 2006.

C. Schalon, J.-S. Surgand, E. Kellenberger, and D. Rognan. A simple and fuzzy method to align and compare druggable ligand-binding sites. *Proteins: Structure, Function, and Bioinformatics*, 71(4):1755–1778, 2008.

J. F. Scheid, H. Mouquet, B. Ueberheide, R. Diskin, F. Klein, T. Y. K. Oliveira, J. Pietzsch, D. Fenyo, A. Abadir, K. Velinzon, A. Hurley, S. Myung, F. Boulad, P. Poignard, D. R. Burton, F. Pereyra, D. D. Ho, B. D. Walker, M. S. Seaman, P. J. Bjorkman, B. T. Chait, and M. C. Nussenzweig. Sequence and Structural Convergence of Broad and Potent HIV Antibodies That Mimic CD4 Binding. *Science*, 333(6049): 1633–1637, 2011.

A. M. Schreyer and T. Blundell. USRCAT: Real-time ultrafast shape recognition with pharmacophoric constraints. *Journal of Cheminformatics*, 4(11):27, 2012.

D. Schritt, S. Li, J. Rozewicki, K. Katoh, K. Yamashita, W. Volkmuth, G. Cavet, and D. M. Standley. Repertoire Builder: high-throughput structural modeling of B and T cell receptors. *Molecular Systems Design & Engineering*, 4:761–768, 2019.

H. W. Schroeder and L. Cavacini. Structure and function of immunoglobulins. *Journal of Allergy and Clinical Immunology*, 125:41–52, 2010.

I. Sela-Culang, V. Kunik, and Y. Ofran. The Structural Basis of Antibody–Antigen Recognition. *Frontiers in Immunology*, 4:302, 2013.

I. Sela-Culang, Y. Ofran, and B. Peters. Antibody specific epitope prediction – emergence of a new paradigm. *Current Opinion in Virology*, 11:98–102, 2015.

A. M. Sevy, J. F. Healey, W. Deng, P. C. Spiegel, S. L. Meeks, and R. Li. Epitope mapping of inhibitory antibodies targeting the C2 domain of coagulation factor VIII by hydrogen–deuterium exchange mass spectrometry. *Journal of Thrombosis and Haemostasis*, 11(12):2128–2136, 2013.

E. Sharon, L. V. Sibener, A. Battle, H. B. Fraser, K. C. Garcia, and J. K. Pritchard. Genetic variation in MHC proteins is associated with T cell receptor expression biases. *Nature Genetics*, 48:995–1002, 2016.

Z. Shi, Q. Zhang, H. Yan, Y. Yang, P. Wang, Y. Zhang, Z. Deng, M. Yu, W. Zhou, Q. Wang, X. Yang, X. Mo, C. Zhang, J. Huang, H. Dai, B. Sun, Y. Zhao, L. Zhang, Y.-G. Yang, and X. Qiu. More than one antibody of individual B cells revealed by single-cell immune profiling. *Cell Discovery*, 5(1):1–13, 2019.

I. N. Shindyalov and P. E. Bourne. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. *Protein Engineering*, 11(9):739–747, 1998.

H. Shirai, A. Kidera, and H. Nakamura. Structural classification of CDR-H3 in antibodies. *FEBS Letters*, 399(1-2):1–8, 1996.

H. Shirai, A. Kidera, and H. Nakamura. H3-rules: identification of CDR-H3 structures in antibodies. *FEBS Letters*, 455(1-2):188–197, 1999.

A. Shulman-Peleg, S. Mintz, R. Nussinov, and H. J. Wolfson. Protein-protein interfaces: Recognition of similar spatial and chemical organizations. In *International Workshop on Algorithms in Bioinformatics*, pages 194–205. Springer, 2004.

A. Shulman-Peleg, R. Nussinov, and H. J. Wolfson. SiteEngines: recognition and comparison of binding sites and protein–protein interfaces. *Nucleic Acids Research*, 33 (Suppl_2):W337–W341, 2005.

I. Sillitoe, T. E. Lewis, A. Cuff, S. Das, P. Ashford, N. L. Dawson, N. Furnham, R. A. Laskowski, D. Lee, J. G. Lees, S. Lehtinen, R. A. Studer, J. Thornton, and C. A. Orengo. CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1):D376–D381, 2015.

A. Sircar, E. T. Kim, and J. J. Gray. RosettaAntibody: Antibody variable region homology modeling server. *Nucleic Acids Research*, 37(Suppl_2):W474–W479, 2009.

S. Sirin, J. R. Apgar, E. M. Bennett, and A. E. Keating. AB-Bind: Antibody binding mutational database for computational affinity predictions. *Protein Science*, 25(2): 393–409, 2016.

C. Soto, R. G. Bombardi, A. Branchizio, N. Kose, P. Matta, A. M. Sevy, R. S. Sinkovits, P. Gilchuk, J. A. Finn, and J. E. Crowe. High frequency of shared clonotypes in human B cell receptor repertoires. *Nature*, 566(7744):398, 2019.

J. W. Stave and K. Lindpaintner. Antibody and Antigen Contact Residues Define Epitope and Paratope Size and Structure. *The Journal of Immunology*, 191(3): 1428–1435, 2013.

C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1): 25, 2007.

E. J. Sutton, R. T. Bradshaw, C. M. Orr, B. Frendéus, G. Larsson, I. Teige, M. S. Cragg, I. Tews, and J. W. Essex. Evaluating Anti-CD32b F (ab) Conformation Using Molecular Dynamics and Small-Angle X-Ray Scattering. *Biophysical Journal*, 115(2): 289–299, 2018.

K. E. Tiller and P. M. Tessier. Advances in Antibody Design. *Annual Review of Biomedical Engineering*, 17(1):191–216, 2015.

I. Trenevska, D. Li, and A. H. Banham. Therapeutic Antibodies against Intracellular Tumor Antigens. *Frontiers in Immunology*, 8:1001, 2017.

A. Tripathi and G. E. Kellogg. A novel and efficient tool for locating and characterizing protein cavities and binding sites. *Proteins: Structure, Function and Bioinformatics*, 78(4):825–842, 2010.

J. Trück, M. N. Ramasamy, J. D. Galson, R. Rance, J. Parkhill, G. Lunter, A. J. Pollard, and D. F. Kelly. Identification of antigen-specific B cell receptor sequences using public repertoire analysis. *The Journal of Immunology*, 194(1):252–261, 2015.

A. C. Wallace, N. Borkakoti, and J. M. Thornton. Tess: A geometric hashing algorithm for deriving 3D coordinate templates for searching structural databases. Application to enzyme active sites. *Protein Science*, 6(11):2308–2323, 1997.

S. Warszawski, A. B. Katz, R. Lipsh, L. Khmelnitsky, G. B. Nissan, G. Javitt, O. Dym, T. Unger, O. Knop, S. Albeck, R. Diskin, D. Fass, M. Sharon, and S. J. Fleishman. Optimizing antibody affinity and stability by the automated design of the variable light-heavy chain interfaces. *PLoS Computational Biology*, 15(8):e1007207, 2019.

B. Webb and A. Sali. Protein structure modeling with MODELLER. In D. Kihara, editor, *Protein Structure Prediction*, pages 1–15. Springer, 2014.

M. Weisel, E. Proschak, and G. Schneider. PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chemistry Central Journal*, 1(1):7, 2007.

B. D. Weitzner, R. L. Dunbrack Jr, and J. J. Gray. The origin of CDR H3 structural diversity. *Structure*, 23(2):302–311, 2015.

B. D. Weitzner, J. R. Jeliazkov, S. Lyskov, N. Marze, D. Kuroda, R. Frick, J. Adolf-Bryfogle, N. Biswas, R. L. Dunbrack, and J. J. Gray. Modeling and docking of antibody structures with Rosetta. *Nature Protocols*, 12(2):401–416, 2017.

W. J. Wilbur. On the PAM matrix model of protein evolution. *Molecular Biology and Evolution*, 2(5):434–447, 1985.

W. K. Wong, G. Georges, F. Ros, S. Kelm, A. P. Lewis, B. Taddese, J. Leem, and C. M. Deane. SCALOP: sequence-based antibody canonical loop structure annotation. *Bioinformatics*, 35(10):1774–1776, 2019a.

W. K. Wong, J. Leem, and C. M. Deane. Comparative analysis of the CDR loops of antigen receptors. *Frontiers in Immunology*, 10:2454, 2019b.

W. K. Wong, C. Marks, J. Leem, A. P. Lewis, J. Shi, and C. M. Deane. TCRBuilder: multi-state T-cell receptor structure prediction. *Bioinformatics*, 36(11):3580–3581, 2020a.

W. K. Wong, S. A. Robinson, A. Bujotzek, G. Georges, A. P. Lewis, J. Shi, J. Snowden, B. Taddese, and C. M. Deane. Ab-Ligity: Identifying sequence-dissimilar antibodies that bind to the same epitope. *BioRxiv*, 2020b.

D. J. Wood, J. d. Vlieg, M. Wagener, and T. Ritschel. Pharmacophore fingerprint-based approach to binding site subpocket similarity and its application to bioisostere replacement. *Journal of Chemical Information and Modeling*, 52(8):2031–2043, 2012.

L. Xie and P. E. Bourne. Detecting evolutionary relationships across existing fold space, using sequence order–independent profile–profile alignments. *Proceedings of the National Academy of Sciences*, 105(14):5441–5446, 2008.

Y. Xu, Z. Yang, L. H. Horan, P. Zhang, L. Liu, B. Zimdahl, S. Green, J. Lu, J. F. Morales, D. M. Barrett, S. A. Grupp, V. W. Chan, H. Liu, and C. Liu. A novel antibody-TCR (AbTCR) platform combines Fab-based antigen recognition with gamma/delta-TCR signaling to facilitate T-cell cytotoxicity with low cytokine release. *Cell Discovery*, 4(1):62, 2018.

Y. Xu, G. T. Salazar, N. Zhang, and Z. An. T-cell receptor mimic (TCRm) antibody therapeutics against intracellular proteins. *Antibody Therapeutics*, 2(1):22–32, 2019.

J. Ye, N. Ma, T. L. Madden, and J. M. Ostell. IgBLAST: an immunoglobulin variable domain sequence analysis tool. *Nucleic Acids Research*, 41:W34–W40, 2013.

Q. Zhang, J. Yang, J. Bautista, A. Badithe, W. Olson, and Y. Liu. Epitope mapping by HDX-MS elucidates the surface coverage of antigens associated with high blocking efficiency of antibodies to birch pollen allergen. *Analytical Chemistry*, 90(19): 11315–11323, 2018.

W. Zhou and H. Yan. Alpha shape and delaunay triangulation in studies of protein-related interactions. *Briefings in Bioinformatics*, 15(1):54–64, 2014.