UNDERSTANDING ANTIBODY BINDING SITES

Jaroslaw Nowak

The Queen's College

University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2017

ABSTRACT

Antibodies are soluble proteins produced by the adaptive immune system to bind and counteract invading pathogens. The binding properties of a typical human antibody are determined by the structure of its variable domain, composed of two chains – heavy and light and by the conformation of six loops located on the surface of the variable domain, known as Complementarity Determining Regions (CDRs).

In the first chapter, we describe our analysis of the conformational space occupied by five out of six antibody CDRs (L1, L2, L3, H1 and H2) and the development of a novel, length-independent method for grouping these CDRs into structural clusters (canonical forms). We show that using our method we can increase coverage and precision of assigning CDR sequences into clusters.

In the next chapter, we describe a method for ranking structural decoys of the CDR-H3 loop. We show that by computationally perturbing CDR-H3 decoys we can improve the performance of existing ranking methods. In the same chapter, we discuss the development of a method for high-throughput assignment of heavy-light chain orientation. The power of the method was demonstrated by assigning orientation to billions of potential Fv sequences.

The third Chapter describes the analysis of a large dataset of CDR sequences with the aim of identifying sequence patterns responsible for the loops' structure. Using a neural network methodology, we found several groups of CDR sequences which might be indicative of previously-unseen conformations.

In the final results Chapter, we describe how we used the structural knowledge developed throughout the rest of the thesis to create a novel pipeline for computational antibody design. We show that the binders developed using our methodology had similar features to available antibody therapeutics and low predicted propensity to cause an immunogenic response. These results demonstrate the potential for using computational methods for designing high affinity therapeutics with human properties.

UNDERSTANDING ANTIBODY BINDING SITES

Jaroslaw Nowak
The Queen's College
University of Oxford

A thesis submitted for the degree of Doctor of Philosophy

Trinity 2017

To Anna and my parents

ACKNOWLEDGEMENTS

First of all, I would like to thank Charlotte for putting up with me for the last four years, and for all the research and life-related advice she shared with me. I couldn't have asked for a better supervisor and I would not have been able to complete this degree without her.

I would also like to thank my other supervisors, Jiye, Seb and Terry at UCB, Guy at Roche, Sid and Boyana at Medimmune, and Alan at GSK. I appreciate all their ideas and help they have given me throughout this project.

I am grateful for having a wonderful research group, OPIG. In particular, I wish to thank the fellow antibody people: Cristian, Claire, Jin, James, Konrad and Alex for all their help and collaboration.

Finally, I would like to thank all my friends and family for the support they provided me in the last four years. Special thanks go to Anna, my partner, for being there for me when I needed it the most.

ABSTRACT

Antibodies are soluble proteins produced by the adaptive immune system to bind and counteract invading pathogens. The binding properties of a typical human antibody are determined by the structure of its variable domain, composed of two chains – heavy and light and by the conformation of six loops located on the surface of the variable domain, known as Complementarity Determining Regions (CDRs).

In the first chapter, we describe our analysis of the conformational space occupied by five out of six antibody CDRs (L1, L2, L3, H1 and H2) and the development of a novel, length-independent method for grouping these CDRs into structural clusters (canonical forms). We show that using our method we can increase coverage and precision of assigning CDR sequences into clusters.

In the next chapter, we describe a method for ranking structural decoys of the CDR-H3 loop. We show that by computationally perturbing CDR-H3 decoys we can improve the performance of existing ranking methods. In the same chapter, we discuss the development of a method for high-throughput assignment of heavy-light chain orientation. The power of the method was demonstrated by assigning orientation to billions of potential Fv sequences.

The third Chapter describes the analysis of a large dataset of CDR sequences with the aim of identifying sequence patterns responsible for the loops' structure. Using a neural network methodology, we found several groups of CDR sequences which might be indicative of previously-unseen conformations.

In the final results Chapter, we describe how we used the structural knowledge developed throughout the rest of the thesis to create a novel pipeline for computational antibody design. We show that the binders developed using our methodology had similar features to available antibody therapeutics and low predicted propensity to cause an immunogenic response. These results demonstrate the potential for using computational methods for designing high affinity therapeutics with human properties.

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.

Jaroslaw Nowak

27th September 2017

PUBLICATIONS

Related to Chapter 2:

Nowak, J., Baker, T., Georges, G., Kelm, S., Klostermann, S., Shi, J., Sridharan, S., and Deane, C. M. 2016. "Length-Independent Structural Similarities Enrich the Antibody CDR Canonical Class Model." *mAbs* 8(4):751–60.

Dunbar, J., Krawczyk, K., Leem, J., Marks, C., **Nowak, J.**, Regep, C., Georges, G., Kelm, S., Popovic, B., and Deane, C. M. 2016. "SAbPred: A Structure-Based Antibody Prediction Server." *Nucleic Acids Research* 44(W1):W474–78.

Related to Chapter 3:

Marks, C., **Nowak, J.**, Klostermann, S., Georges, G., Dunbar, J., Shi, J., Kelm, S., and Deane, C. M. 2017. "Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve Protein Loop Prediction." *Bioinformatics* 33(9):1346–53.

Parks, T., Mirabel, M. M., Kado, J., Auckland, K., **Nowak, J.**, Rautanen, A., Mentzer, A. J., Marijon, E., Jouven, X., Perman, M. L., et al. 2017. "Association between a Common Immunoglobulin Heavy Chain Allele and Rheumatic Heart Disease Risk in Oceania." *Nature Communications* 8:14946.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

LIST OF ABBREVIATIONS

AID – Activation-Induced Cytidine Deaminase

AMA – Antibody Modelling Assessment

AML – Antibody Model Library

ANN – Artificial Neural Network

AUC – Area Under the Curve

CA – Carbon Alpha

CB – Carbon Beta

CCD – Cyclic Coordinate Descent

CDR – Complementarity Determining Region

CH – Constant Heavy

CL – Constant Light

DBSCAN – Density-Based Spatial Clustering of Applications with Noise

DNA – Deoxyribonucleic Acid

DTW – Dynamic Time Warping

ESST – Environment Specific Substitution Table

FFT – Fast Fourier Transform

FPR – False Positive Rate

HMM – Hidden Markov Model

KIC – Kinematic Closure

MHC – Major Histocompatibility Complex

MLR – Machine-Learned Ranking

NGS – Next Generation Sequencing

NGS – Next-Generation Sequencing

OPTICS – Ordering Points to Identify the Clustering Structure

PCA – Principal Component Analysis

PDB – Protein Data Bank

PSSM – Position Specific Scoring Matrix

PSST – Position Specific Scoring Table

RMSD – Root Mean Square Deviation

ROC – Receiver Operating Characteristic

SAbDab – Structural Antibody Database

SASA – Solvent Accessible Surface Area

TCR – T-Cell Receptor

TM-score – Template Modelling score

TPR – True Positive Rate

UPGMA – Unweighted Pair Group Method with Arithmetic mean

VH – Variable Heavy

VL – Variable Light

1 INTRODUCTION

1.1 Introduction

Vertebrate antibodies are proteins produced by the immune system to detect and act upon foreign objects (antigens). Their specificity and high affinity makes lab-produced antibodies attractive for various research applications such as flow cytometry, immunoprecipitation, ELISA or Western blot. Monoclonal antibodies (mAbs) are also used as therapeutics. In 1986 the first monoclonal antibody drug (Muromonab-CD3) was approved by the FDA.

The aim of this thesis is to analyse antibody structure and sequence data, with the goal of creating a computational pipeline for designing high-affinity protein therapeutics. The first two results chapters of this work contain a study of antibody's structural elements and different algorithms for antibody structure prediction. The third chapter is an analysis of a large Next-Generation Sequencing (NGS) dataset of antibodies. The last results chapter combines the concepts introduced throughout the rest of the thesis to create a methodology for computational antibody design.

This introductory chapter provides theoretical background to the research presented throughout the thesis. First, we briefly describe factors responsible for protein structure and function. Next, we introduce the biology of the human immune system with an emphasis on antibody function. Following that, we present available methods for antibody structure prediction and their use in computational antibody design. Finally, we summarise the contents of following chapters of this thesis.

1.2 Protein structure

Proteins are small biological machines. Virtually every activity of a living organism is performed using proteins. The sequence (and therefore structure) of a protein is coded for by Deoxyribonucleic Acid (DNA) sequence. To form a protein the DNA sequence is first transcribed onto messenger Ribonucleic Acid (mRNA) and then translated onto a polypeptide chain of amino acids. The amino acid chain then self-assembles into a three-dimensional protein though interactions between amino acids.

1.2.1 Amino acids

Proteins are composed of smaller units known as amino acids. An amino acid is composed of four backbone atoms and a side chain, known also as an R group. The four backbone atoms are: carbonyl oxygen (O), carbonyl carbon (C), alpha carbon (CA) and the nitrogen (N). The functional group R is attached to the alpha carbon and gives an amino acid distinct chemical and physical properties.

There are 20 canonical amino acids (see Figure 1.1). There are many ways of grouping the amino acids, in Figure 1.1 they are clustered by their physical and chemical properties and by charge, determined by the respective side chains.

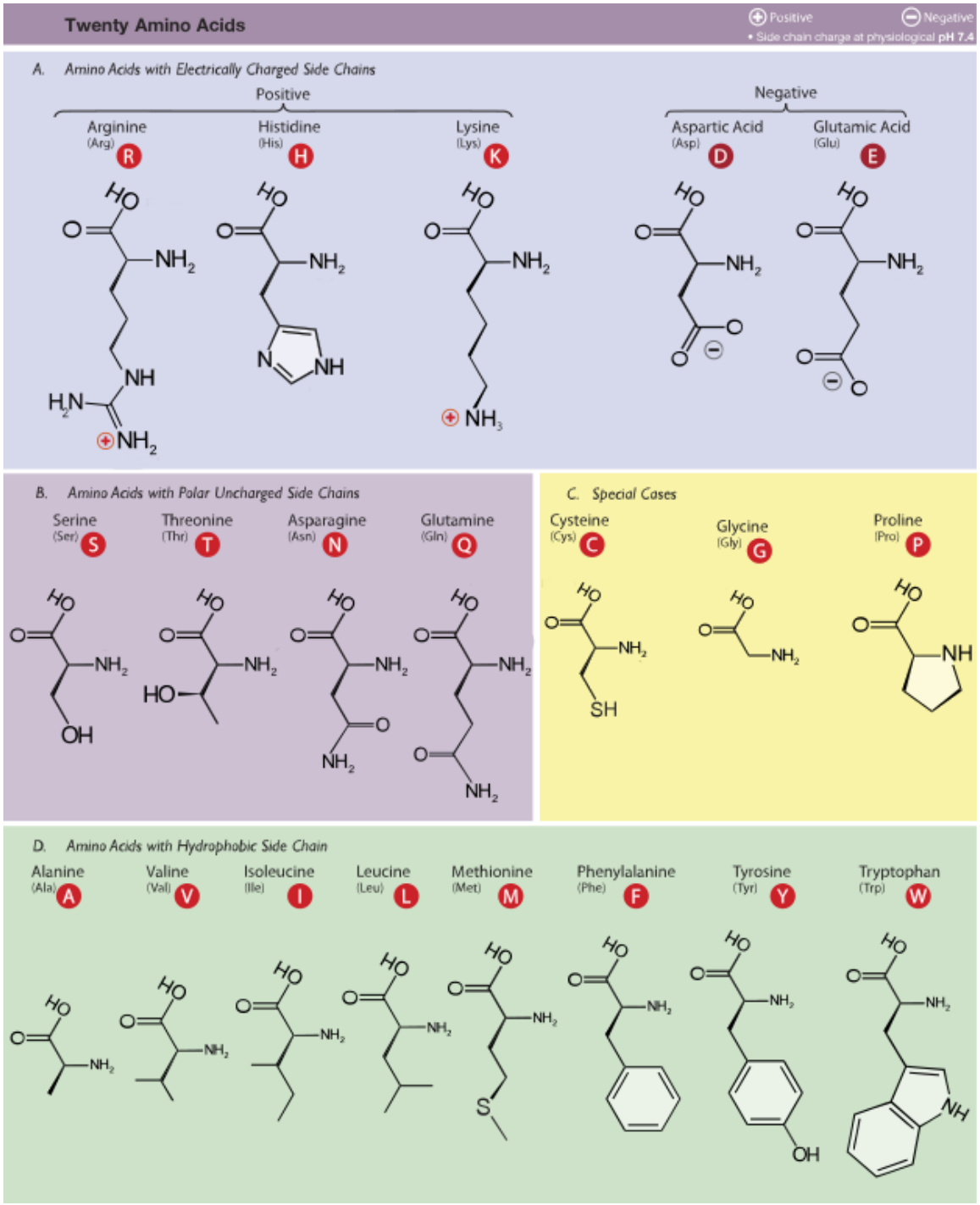

Figure 1.1 The twenty proteinogenic amino acids. The schematic shows the chemical composition of the building blocks of mammalian proteins. The amino acids are grouped by their chemical and physical properties (Charged, Polar, Hydrophobic and Special) and by their charge (Positive and Negative) at pH 7.4. The figure was adapted from the image created by Dancojocari

1.2.2 Hierarchy of protein structure elements

In a protein structure, amino acids are joined together through covalent peptide bonds, formed between the backbone carbonyl carbon of one amino acid and the backbone nitrogen of another. The polypeptide chain of amino acid units, called residues, is known as the primary structure of a protein. By convention, the polypeptide chain begins at the backbone nitrogen of the first residue (known as N-terminus) and ends at the backbone carbonyl carbon of the last residue (known as the C-terminus). A typical protein polypeptide chain consists of between 50 and 2,000 amino acids (Berg et al. 2012).

In 1951 Linus Pauling and Robert Corey noticed that polypeptide chains tend to assemble into repeating structural segments, called α-helices and β-sheets (Pauling et al. 1951). These segments, known as the secondary structure of the protein, are maintained by hydrogen bond networks between backbone atoms of adjacent residues (see Figure 1.2). The α-helix if formed by hydrogen bonds between the CO group of a residue and the NH group of a residue four positions further in the sequence.

β-sheets are composed of polypeptide chains, called β-strands, held together through hydrogen bonds between CO groups and NH groups of adjacent residues. β-sheets occur in two variants – an antiparallel β-sheet, where each residue on each β-strand is bonded to two residues on the adjacent sheets and a parallel β-sheet, where each residue on each β-strand is bonded to one residue on the adjacent sheet. The β-sheet shown in Figure 1.2 is an example of a parallel β-sheet.

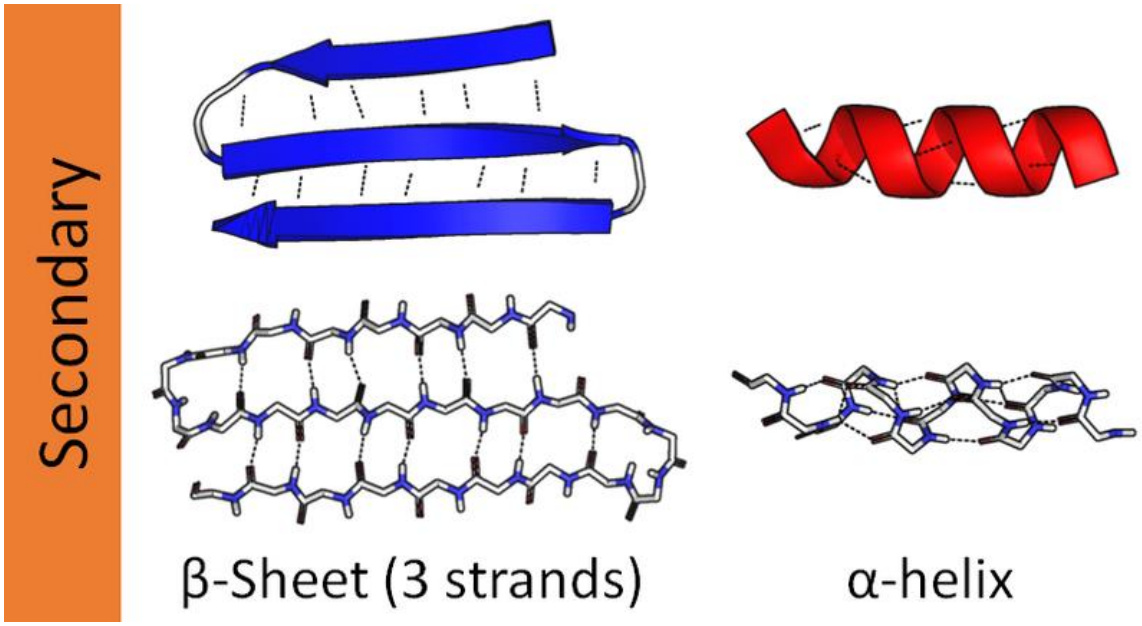

Figure 1.2 Protein secondary structure. The schematic shows the hydrogen bond networks responsible for the two secondary structure elements – the β-sheet and the α-helix. The figure was created by Thomas Shafee and was reproduced from https://commons.wikimedia.org/ under CC BY-SA 4.0 licence (https://creativecommons.org/licenses/by-sa/4.0/deed.en).

Not every amino acid is part of a well-defined secondary structure segment. Segments not designated as secondary structure are known as loops. They connect secondary structure elements (unless the protein does not form any secondary structure) and often lie on the surface of a protein. This positioning makes them important for protein-protein interactions. Despite not forming regular patterns, loops can be rigid (Marks et al. 2016).

Secondary structure elements fold into a three-dimensional construct, known as the tertiary structure of the protein. A typical soluble protein consists of a hydrophobic core, composed of non-polar amino acids and of a hydrophilic surface, composed of charged and polar residues. Some proteins consist of several independent domains, containing 30 to 400 amino acids, connected to each other through polypeptide chains.

The individual polypeptide chains often assemble into a larger construct, known as the quaternary structure of the protein, which is the highest level of protein structure organization. The individual units in a quaternary structure are not connected to each other through peptide bonds, yet they form a stable protein. Often the individual polypeptide chains which are part of such a protein cannot exist in isolation and may unfold if separated.

The structure of a protein can be described using three dihedral (or torsion) angles of protein backbone (Richardson 1981). These are known as the φ (phi), ψ (psi) and ω (omega). The calculation of these torsion angles involves the following backbone atoms, respectively (see Figure 1.3):

- φ: C_{-1}, N, CA, C
- ψ: N, CA, C, N_{+1}
- ω: CA_{-1}, C_{-1}, N, CA

Where CA is the carbon alpha, C is the carbonyl carbon, O is the backbone oxygen and N is the backbone nitrogen. The subscripts -1 and +1 indicate the preceding and succeeding amino acids respectively.

The omega dihedral angle can take on only two values: 0° (cis configuration) and 180° (trans configuration). Phi and psi angles are confined to particular areas of coordinate space (Ramachandran et al. 1963). A plot visualizing the allowed areas of phi-psi angles is known as the Ramachandran plot (Ramachandran et al. 1963) (see Figure 1.3)

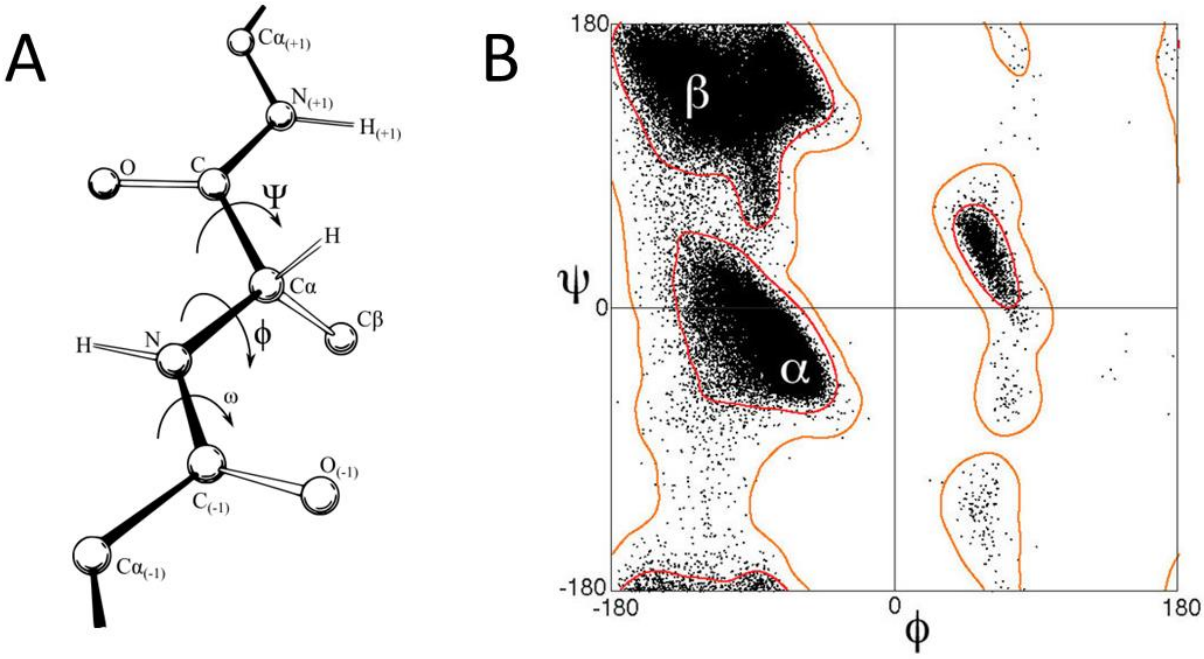

Figure 1.3 The dihedral angles. Panel A shows the atoms involved in the definition of each dihedral angle. CA is the Carbon Alpha, C is the carbonyl Carbon, N is the backbone Nitrogen, O is the backbone Oxygen and CB is the Carbon Beta. The subscripts -1 and +1 refer to the preceding and succeeding amino acid respectively. Panel B shows the areas of the coordinate space occupied by the allowed pairs of phi, psi angles. The dots represent amino acid dihedral angles calculated by (Lovell et al. 2003). The areas occupied by amino acids which are part of α helices or β sheets are highlighted. The density plot shown in Panel B is known as the Ramachandran plot (Ramachandran et al. 1963). Both panels were created by Wikimedia Commons user Dcrjsr and were reproduced under CC BY 3.0 licence (https://creativecommons.org/licenses/by/3.0/deed.en).

1.2.3 Protein structure determination

In this section, we describe how a three-dimensional structure of a protein is determined experimentally. Such structural characterisation can be a source of valuable knowledge, such as the proteins functionality or specificity of its binding sites. The Protein Data Bank

(PDB, https://www.rcsb.org/) (Berman et al. 2000) is an online database which stores the three-dimensional protein structures and makes them available to researchers free of charge. As of August 2017, the PDB contains over 123,000 protein structures.

Protein structure is usually determined using X-ray crystallography (Kendrew et al. 1958). Over 90% of structures available in the PDB were solved using this method. To conduct an X-ray experiment, the protein of interest must first be crystallized. In a crystal, individual protein components are arranged into repeating units through amino acid interactions. Crystallization can often be difficult and often hundreds of trials must be conducted to find optimum conditions. When a suitable crystal is obtained, it is exposed to an X-ray beam which scatters off the component proteins. The resulting scattering pattern is then used to determine the three-dimensional coordinates of the atoms forming the target protein.

In X-ray crystallography, alongside three-dimensional coordinates, two additional features are recorded for each atom. The first is called the temperature factor, also known as the B-factor (Berman et al. 2000). The B-factor describes the attenuation of X-rays caused by random thermal motion and is a measure of uncertainty in the atom position. It is measured in square Angstroms ($Å^2$). The second feature is the occupancy, which measures the fraction of crystal units in which a given atom is in a given position (Berman et al. 2000). As some protein fragments, can assume more than one conformation, this field gives the information about their relative frequency. The overall quality of X-ray structures is also indicated by two additional metrics – the resolution and the R-factor. Resolution describes the minimum distance required for two atoms to be resolved. A resolution below 1Å indicates excellent quality of data, resolution between 1-3.5Å indicates good quality data and resolutions above 3.5Å usually indicate poor

quality of data (Brunger et al. 2009). The R-factor quantifies how well does the structural model explain the observed crystallographic data. An R-factor of 0.0 indicates a perfect model, while an R-factor above 0.63 indicates a random model (Kleywegt et al. 1997).

The quality measures described above are commonly used to filter out low-quality structures in a large-scale analysis (e.g. North et al. 2011).

1.2.4 Protein structure prediction

The methods described in the previous section are costly and time consuming, which limits the number of structurally characterised proteins (Brunger et al. 2009). As mentioned before, at the time of writing this thesis, there are over 123,000 publicly available protein structures. At the same time, the UniProt database (The UniProt Consortium 2017) contains over 80,000,000 protein sequences, highlighting that only a small fraction of known protein sequences have been structurally characterised.

The field of protein structure prediction aims to develop a systematic method for inferring protein structure from sequence, with little to no experimental input. This is usually accomplished by creating an assembly of trial structural models, called decoys, and estimating their energy using a combination of statistical and physics-based approaches. The lowest energy model should resemble the true protein structure. Existing structure prediction algorithms can be broadly categorised into template-based methods (also known as knowledge-based or homology-based methods) and template-free methods (also known as *de novo* or *ab initio* methods) (Walsh et al. 2009).

The accuracy of a given modelling technique can be benchmarked by modelling structurally characterised proteins and comparing coordinates of selected atoms between model and the true structure. The atoms selected for comparison are usually Cα or backbone atoms of equivalent residues. The simplest technique for comparing atomic coordinates involves aligning the proteins using Kabsch algorithm (Kabsch 1976) and calculating Root Mean Square Deviation (RMSD) over equivalent atoms according to the following formula:

$$RMSD = \sqrt{\frac{\sum_{i=0}^{N}(x_{i,M} - x_{i,S})^2 + (y_{i,M} - y_{i,S})^2 + (z_{i,M} - z_{i,S})^2}{N}}$$

Where x, y, z are the cartesian coordinates of the constituent atoms, i is the index of a given atom, N is the total number of atoms and S, M subscripts indicate Structure and Model respectively.

Models are usually considered accurate if they are within 3Å of the true structure, which is the expected deviation between highly homologous proteins (Chothia et al. 1986). A stricter threshold of about 1.5Å is often considered when modelling a smaller fragment of a protein, such as a single loop (Marks et al. 2017). RMSD is length-dependent (Carugo et al. 2001), making it hard to compare accuracies across very different protein sizes. Several other accuracy measures have been developed, for example the Template Modelling score (TM-score) (Zhang et al. 2004). The TM-score is calculated between two protein structures using the following sigmoidal formula:

$$TMscore = \max\left[\frac{1}{L_N}\sum_{i=1}^{L_r}\frac{1}{1 + \left(\frac{d_i}{d_0}\right)^2}\right]$$

Where the $\max$ function indicates the optimal alignment, producing the highest score, L_N is the length of the target protein, L_r is the length of the aligned region, d_i is the Euclidean distance between ith pair of residues in the alignment and d_0 is the sigmoidal mid-point, defined such that the TM-score is length independent. The d_0 parameter was heuristically defined to be $d_0 = 1.24\sqrt[3]{L_N - 15} - 1.8$ (Zhang et al. 2004). Due to its sigmoidal construction, the TM-score is constrained between 0 and 1. A structural model is typically considered accurate when the TM-score to the true crystal structure is above 0.5 (Xu et al. 2010). A TM-score below 0.2 indicates a random model (Zhang et al. 2005).

Despite significant advances in recent decades, protein structure prediction remains an open problem. Accuracy of existing protein structure prediction algorithms is benchmarked every two years in a community-wide CASP (Critical Assessment of protein Structure Prediction) competition (Moult et al. 2014).

1.3 Antibody biology

This section focuses on one type of protein – the antibody. Here, we give a biological overview of how antibodies are produced, their structure and functionality. We also discuss how lab-produced antibodies are used as therapeutics. Finally, we review the available sources of antibody sequence and structure data.

1.3.1 The immune system

The immune system of an organism protects it against disease. It recognizes a wide variety of pathogens (e.g. fungi, bacteria or viruses) and neutralizes them. The immune system also possesses a remarkable ability to distinguish invading pathogens from

healthy tissue. It consists of two distinct parts: the innate immune system, present in most living organisms and the adaptive immune system, found only in vertebrates.

The innate immune system is the first line of defence, responding immediately to infection. It consists of elements such as epithelial barriers (physical shells that are hard to penetrate) or phagocytic leukocytes (white cells that digest the pathogens). The components of the innate immune system are immutable in structure and sequence. Therefore it is relatively easy for the pathogens to evolve defences against the innate system (Berg et al. 2012).

If the innate system gets overpowered by the infection, the organism of a vertebrate activates its adaptive immune system. The adaptive system consists of two main response modes – the cell mediated response and the humoral immune response. The cell mediated response utilises a special type of cells called cytotoxic T lymphocytes (or killer T cells) to destroy cells that have been compromised by the infection. The humoral immune response involves production of antibodies by the plasma cells, derived from B lymphocytes, or B-cells (Chaffey et al. 2003).

1.3.2 Antibody function

Antibodies are proteins produced by the immune system that detect and neutralise foreign objects which enter the organism. Typical antibodies are Y-shaped and are composed of two chains – heavy and light. The chain nomenclature is a consequence of the relative weight of the two chains – the molecular weight of the heavy chain is roughly 50 kDa (kilodaltons) while the weight of the light chain is roughly 25 kDa. We describe the antibody structure in more detail in Section 1.4.1. An antibody's functionality is mostly conveyed through binding to invader molecules, known as antigens. The area of

an antibody which binds the antigen is known as a paratope while the part of an antigen the antibody is complementary to is called an epitope.

There are three main avenues for antibodies fighting invading pathogens. These are known as neutralisation, opsonisation and complement activation (Murphy et al. 2012).

Pathogens such as intracellular bacteria or viruses contain proteins on their surface that allow them to bind to the surface of a target cell and eventually get inside the cell. Other pathogens release toxins harmful to the organism. Neutralization occurs by antibodies directly binding to the pathogens or toxins, preventing them from attaching to their intended target (Murphy et al. 2012). The pathogens coated in antibodies also tend to clump together in a process known as agglutination, impairing their mobility. The invader is eventually digested by phagocytes.

In opsonisation, the antibodies do not neutralize the pathogen directly, but instead mark it for destruction by other elements of the immune system (Murphy et al. 2012). The immune system accomplishes this by coating the pathogen in antibodies complementary to its surface proteins. The Fc region of these antibodies (see Section 1.4.1) is complementary to receptors on a phagocyte which envelops and digests the invader. The peptide fragments of the digested pathogen can be used by the phagocyte to further activate other parts of the immune system.

In complement activation, the antibodies recruit innate immune system proteins to bind the pathogen and mark it for destruction. The complement proteins might mark the pathogen for phagocytosis or directly destroy the invader by damaging the membrane and causing lysis (Murphy et al. 2012).

1.3.3 Antibody ontogeny and maturation

There are five types of antibodies, IgM, IgA, IgD, IgE and IgG. They differ in the structure of their hinge region (see Figure 1.4) and function they perform. Here, we describe how the different antibody isotypes are generated during immune response and what function they perform.

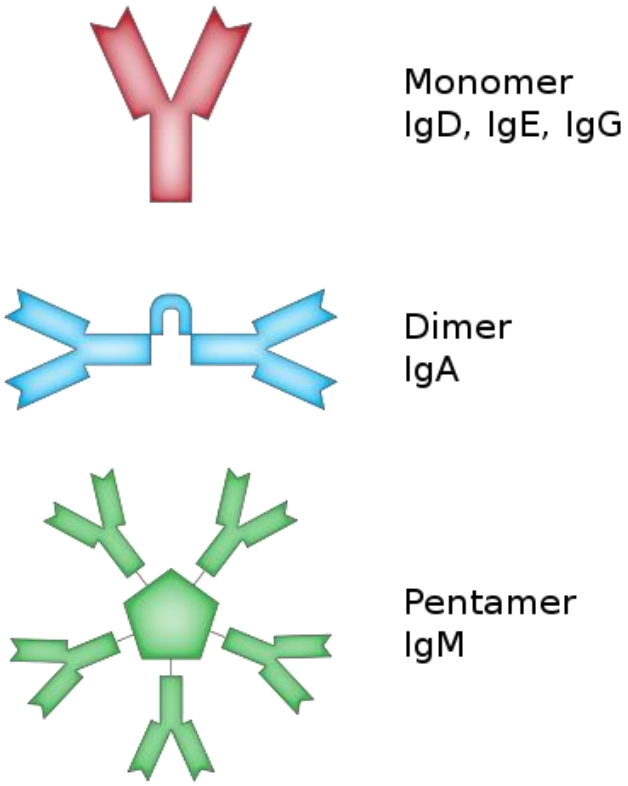

Figure 1.4 The five antibody isotypes. The schematic shows a cartoon representation of the structure of five antibody isotypes. The three monomeric antibody isotypes (IgG, IgE, IgD) are shown in red, the dimeric IgA is shown in blue and the pentameric IgM is shown in green. The image was created by Martin Brändli and was reproduced under CC BY-SA 2.5 licence (https://creativecommons.org/licenses/by-sa/2.5/deed.en).

B lymphocytes differentiate from hematopoietic stem cells in the bone marrow. At the initial stages of B-cell development the lymphocytes only produce receptors of IgM

isotype, containing heavy chains bound to surrogate light chains (Murphy et al. 2012). If the B cell successfully synthesises a light chain, a full IgM receptor is expressed on its surface (see Figure 1.5). The newly produced B-cells undergo a negative selection where they are destroyed through apoptosis if they bind any self-molecules (molecules normally produced by the organism). If they pass this selection they are relocated to the spleen where they are further differentiated into naïve B-cells, capable of producing antibodies of IgM and IgD isotype. At this stage, the B-cells are called naïve because they are not specific for any given antigen. From the spleen, the naïve B-cells are distributed throughout the lymph nodes (Murphy et al. 2012).

The innate immune system contains a number of phagocytic cells which act as messengers between the innate immune system and the adaptive immune system. They phagocytise pathogens, display their fragments on the surface and travel to the lymph nodes to present the fragments to members of the adaptive system.

In the lymph nodes, there are many antibody-producing B-cells with diverse binding sites. When a phagocyte enters the lymph node, carrying pathogen fragments, there will usually be some B-cells capable of binding the pathogen fragments (Harwood et al. 2010). Once a B-cell binds the antigen, the antigen is internalized into the B-cell and chopped into fragments which are then displayed on the cell's surface using the Major Histocompatibility Complex class II (MHC-II). The displayed peptides are recognized by helper T-cells, activated by the same antigens. The helper T-cells bind the MHC-II using T-Cell Receptors (TCRs) and release cytokines stimulating B-cells to proliferate. Some of the B-cells differentiate into short-lived plasmablasts which release low-affinity antibodies while others move to newly-created germinal centres (Murphy et al. 2012). In the germinal centres, the antibody-coding DNA fragments undergo intensive

modifications in order to increase the antibodies' affinity to the antigen (Shlomchik et al. 2012) (see Section 1.3.4.2). It is also where immunoglobulin class-switching occurs, generating the remaining antibody isotypes (IgA, IgE, IgG). During the maturation process B-cells transform into plasma cells, produced to fight the immediate infection and to memory B-cells which provide long-term protection against recurring incursion (Murphy et al. 2012). From there the cells migrate depending on their isotype. IgAs occur mainly in tissues generating secretions, such as breasts or salivary glands, IgEs are found in epithelial areas such as skin or respiratory tract, while IgG, the most common antibody type, is found throughout the organism in circulating blood and tissues (Murphy et al. 2012).

Once the infection has been cleared, the memory B-cells remain in the organism, ready to be activated if the complementary antigen is encountered again.

1.3.4 Antibody diversity generation.

The humoral immune response would not be possible if not for the extraordinary diversity of antibody producing cells. It has been estimated that a healthy human organism is capable of producing over 10^8 antibodies with distinct sequences (Glanville et al. 2009). Here, we describe the mechanisms responsible for this diversity.

1.3.4.1 V(D)J recombination

The DNA strand coding for an antibody is assembled from three types of gene segments - the Variable (V), Diversity (D), Joining (J) and Constant (C) segments. They are known as the germline genes or germlines. The region of the antibody which contains the antigen-binding site is known as the variable domain fragment or Fv (Zdanov et al. 1994). The Fv region is composed of two chains – heavy and light.

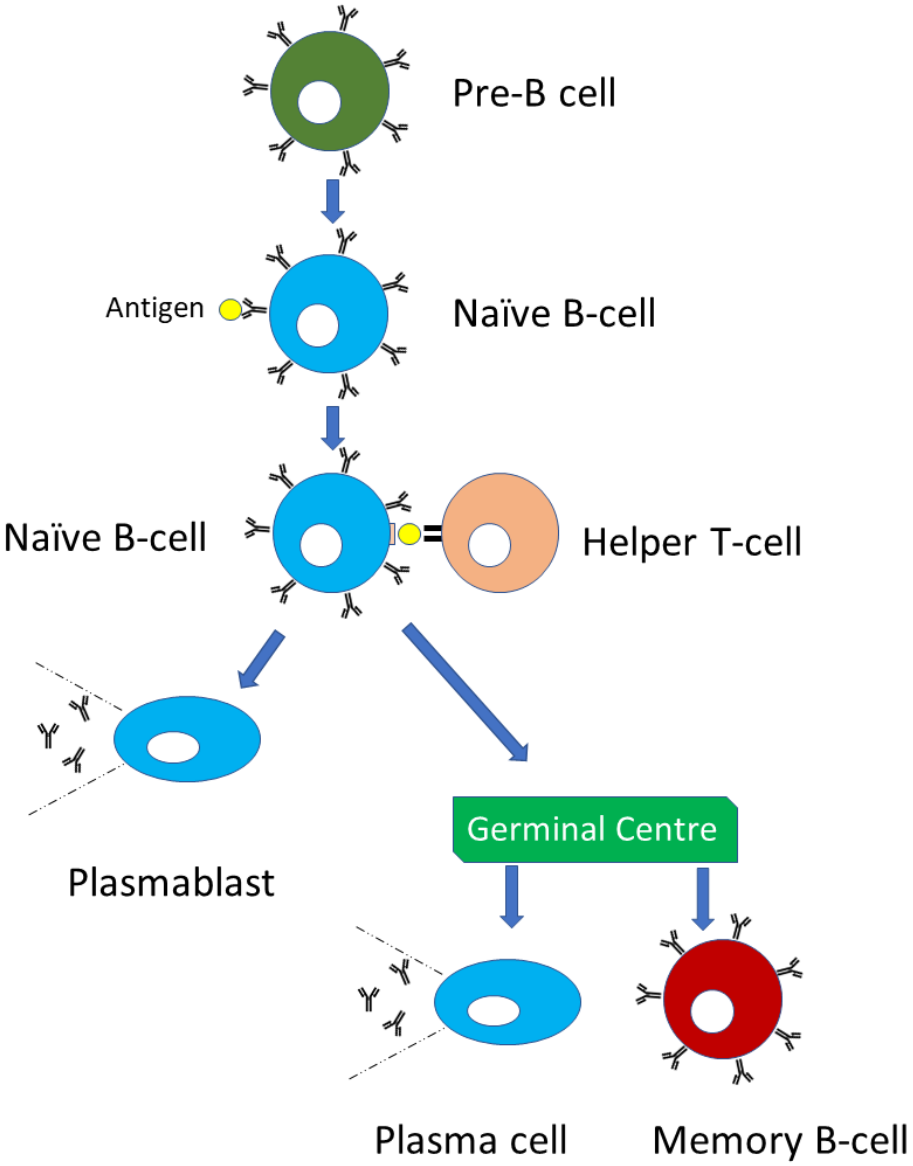

Figure 1.5 Antibody generation. The schematic shows the typical development of a B-cell from the initial pre-B cell stage (shown in green) to plasma cells and memory B-cells (shown in blue and red respectively). First, the pre-B cells differentiate into naïve B-cells. Upon antigen stimulation, some of naïve B-cells differentiate into short-lived plasmablasts which provide an initial immune response. Other stimulated B-cells migrate to a newly-created germinal centre where they undergo affinity maturation and differentiate into plasma cells, which are responsible for fighting the current infection, and into memory B-cells which remain in the organism in case of recurring infection

The light chain is coded for by a light variable gene and by a light joining gene and typically contains about 111 amino acids, out of which ~98 are coded for by the V-gene and ~13 are coded for by the J-gene (Murphy et al. 2012). There are two types of variable light genes present in the human genome – kappa (κ) and lambda (λ).

The heavy chain is, on average, longer than the light and contains 111 to 125 amino acids (Galitsky et al. 1998) out of which the additional ~14 amino acids are added by the D-segment (inserted between the V-segment and the J-segment) (Murphy et al. 2012).

The human genome contains 40 light V genes of type κ, 30 light V genes of type λ, 65 heavy V genes, 27 D genes, 5 light J genes of type κ, 4 light J genes of type λ and 6 heavy J genes (Murphy et al. 2012). When naïve B-cells are generated in the bone marrow (see Section 1.3.3) these genes are spliced together (separately for each chain type) in a process known as the V(D)J recombination (see Figure 1.6). It has been shown (Jackson et al. 2013) that this splicing is not random and some germlines are preferentially used in the recombination process. It has been hypothesised that these biases have arisen through evolution to increase the prevalence of antibodies with certain specificities (Jackson et al. 2013).

Without considering the junctional diversity, the V(D)J recombination can theoretically generate (40*5 + 30*4) * (65*27*6) = 3.4 * 10^6 unique Fv sequences. This figure is five orders of magnitude less than the true diversity.

1.3.4.2 Somatic hypermutation

When B-cells mature in the germinal centres (see Section 1.3.3) they undergo a process known as somatic hypermutation.

During somatic hypermutation the DNA strand coding for the antibody accumulates mutations at a rate of ~10^{-3} mutations/basepair/generation which is about 10^3 times higher than that observed for non-antibody genes (Teng et al. 2007). The mutations are induced by an enzyme called Activation-Induced cytidine Deaminase (AID) which randomly deaminates cytosine nucleotides converting them to uracil creating guanine – uracil pairs.

When DNA is replicated the mismatch-containing strand can diverge into two daughter strands, the first being identical to the unmutated strand and the second containing a thymine – adenine pair at the mismatch site (thymine being the DNA equivalent of uracil). Alternatively, the mismatch site can be repaired in an error-prone way, introducing mutations in the strand.

Following a number of mutations, B-cell producing the modified antibody undergoes selection – if the modified sequence codes for a higher affinity antibody the cell is stimulated to proliferate, otherwise it undergoes apoptosis (Murphy et al. 2012).

Somatic hypermutation increases the antibody diversity by ~10^3 (Murphy et al. 2012). Together with V(D)J recombination and junctional diversity these processes are responsible for the breadth of the antibody repertoire.

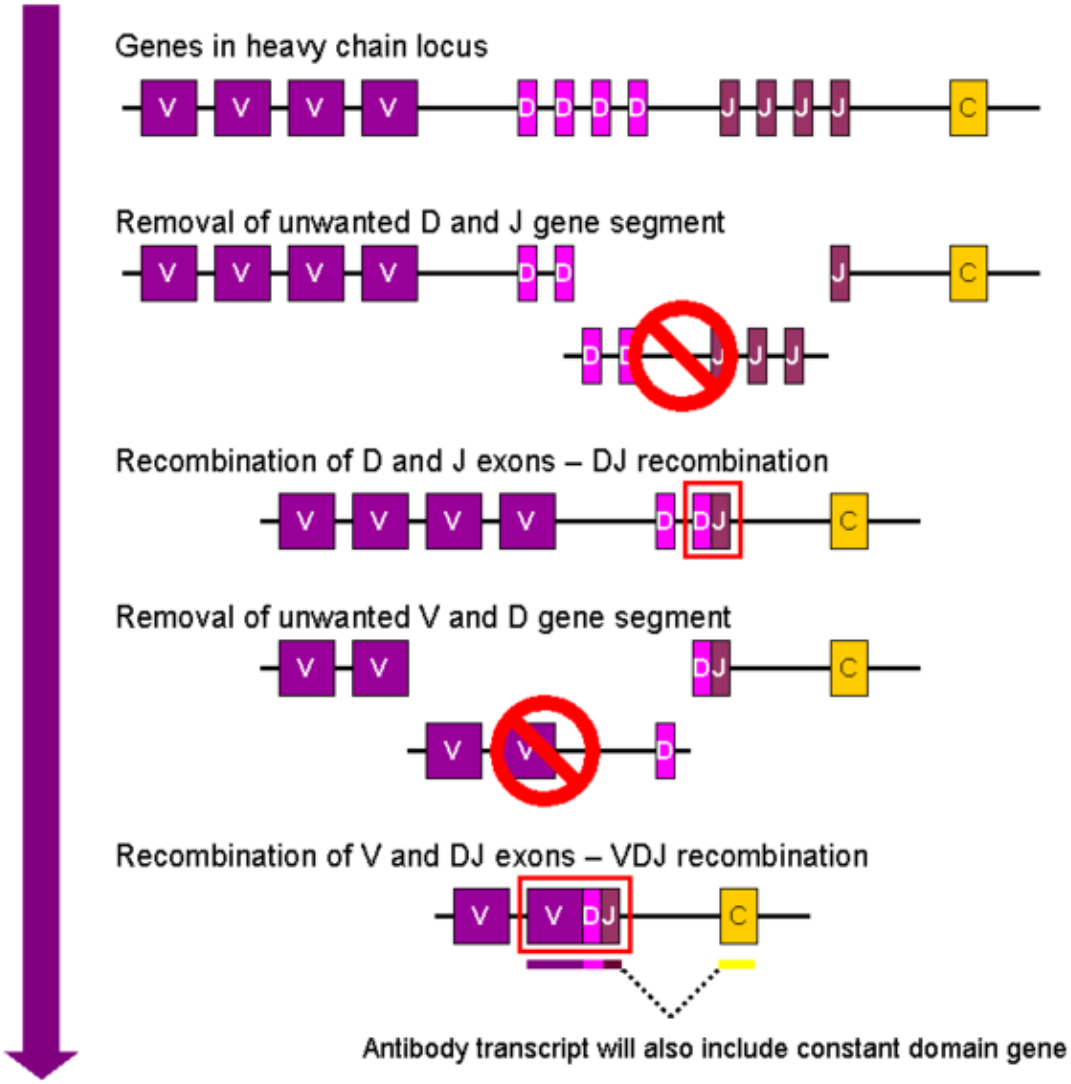

Figure 1.6 The VDJ recombination. The schematic summarises gene splicing process responsible for generation of an antibody heavy chain sequence. First, the D and J gene segments are brought together, deleting any intermediate DNA. The V segment and the DJ segment are brought together in the same fashion. Finally, the combined VDJ sequence is attached to an antibody constant gene. The figure was created by gustavocarra and was reproduced under public domain.

1.3.5 Thermodynamics of antibody binding

As discussed before, the functionality of antibodies is conveyed through their interactions with other molecules. The strength of the interaction is typically quantified

through the dissociation constant, also known as affinity, expressed as the ratio of the concentrations of unbound antibody and antigen to the concentration of the complex. The dissociation constant can be calculated using the following formula:

$$K_D = \frac{[Ab][Ag]}{[AbAg]}$$

Where K_D is the dissociation constant, $[Ab]$ is the concentration of the antibody, $[Ag]$ is the concentration of the antigen and $[AbAg]$ is the concentration of the antibody-antigen complex. Stronger interactions lead to an increase in the denominator and decrease in the numerator resulting in lower values of the dissociation constant.

For binding to occur the complex must have lower Gibbs free energy than the dissociated antibody and antigen (Olsson et al. 2008). The change in Gibbs free energy is related to the dissociation constant through the following formula:

$$\Delta G = -RTlog(K_D)$$

Where ΔG is the change in Gibbs free energy, R is the ideal gas constant (8.3144598(48) J mol^{-1} K^{-1}) and T is the temperature, usually taken to be 300 K.

The change in Gibbs free energy can also be calculated in terms of change in enthalpy (also known as internal energy) and entropy using the following formula:

$$\Delta G = \Delta H - T\Delta S$$

Where ΔH is the change in enthalpy upon binding, T is the temperature and ΔS is the change in entropy upon binding. The change in enthalpy ΔH is the amount of heat released during binding. ΔH is related to formation of residue-residue bonds during binding, such as salt bridges, hydrogen bonds, van der Waals interactions etc. The change is entropy ΔS corresponds to the decrease in the degree of disorder in the system.

Negative values of ΔS are typically related to changes in solvation of hydrophobic residues, while positive values of ΔS are typically related to decrease in the conformational disorder of the residues at antibody-antigen interface (Olsson et al. 2008).

As mentioned before, during the antibody maturation process, the affinity of an antibody towards its antigen is increased by several orders of magnitude (see Section 1.3.3). The enthalpic component of the Gibbs free energy of binding is increased through point mutations at the antibody-antigen interface, creating new chemical bonds between the antibody and antigen (Teng et al. 2007). The entropic component is decreased by increasing the hydrophobic complementarity between antibody and antigen, and by mutations that increase the stability of the antibody structure. The rigidification of the antibody is beneficial, as it minimizes the change in the degree of disorder upon binding (Manivel et al. 2000). The structural elements that typically display the highest degree of decrease in disorder are the CDR loops, especially CDR-H3, (Sela-Culang et al. 2013) and the VH-VL orientation (Dunbar et al. 2013).

The rigidification of the CDR loops gives them unique properties. In general proteins, loops tend to be the most flexible part of the structure (Nilmeier et al. 2011). In contrast, the CDRs have often been observed to be less flexible than the rest of the antibody structure, as quantified by them having lower average temperature factors than the antibody framework (Regep et al. 2017). The existence of the antibody canonical classes (see Section 1.4.3) is also an indirect consequence of the rigidification process, as mutations changing the structure of a canonical CDR loop are likely to destabilize it and therefore be selected against (Teng et al. 2007).

1.3.6 Antibody therapeutics

The specificity and high affinity of antibodies makes them attractive as drug candidates. The first methodology for developing antibodies for a specific target in a lab involved immunizing a model organism (originally a mouse) with the target protein and extracting the matured B-cells (Kennett 1979). The extracted B-cells were immortalized using the hybridoma technology (Köhler et al. 1975) and cloned to produce monoclonal antibodies (mAbs). The disadvantage of the hybridoma technology is that structure of a mouse antibody can be sufficiently different from the structure of a human antibody that, if a mouse antibody is injected into a human, an immune reaction against the drug might occur. To produce mAbs with minimal immunogenicity the non-human antibody needs to be humanized. Humanization is a process which increases the sequence and structure similarity to antibodies produced by humans, while retaining the binding properties of the parental molecule (Almagro et al. 2008). Humanization involves replacing parts of target antibody sequence with elements of human antibody sequence. Humanization methods include chimerization (Morrison et al. 1984), CDR grafting (Queen et al. 1989) and SDR grafting (Kashmiri et al. 2005).

Since the discovery of the first protocol for mAb development there has been a drive to create a methodology for generation of fully human antibodies. Two commonly used approaches are phage display (McCafferty et al. 1990) and the use of transgenic mice (Green et al. 1994; Lonberg et al. 1994). Phage display is an in vitro methodology, where the antibodies are expressed on the surface of a bacteriophage virus. Viruses that express antibodies with increased affinity are then selected (McCafferty et al. 1990). The transgenic mice technology involves creating mice that have been engineered to express antibodies coded for by human germline genes. The antibody-producing B-cells are then obtained from the transgenic mice in a manner similar to the original model

organism methodology described above (Green et al. 1994; Lonberg et al. 1994). At the end of the process, matured B-cells, expressing antibodies with human properties, are extracted.

These methodologies, and others, have resulted in the discovery of many successful antibody therapeutics. The first monoclonal antibody drug (Muromonab-CD3) was approved by the FDA in 1986. Since then the global market for therapeutic mAbs has grown exponentially - in 2010 the combined sales of 25 actively marketed mAbs generated around $43 billion (Elvin et al. 2013). Monoclonal antibody drugs currently available on the market include Humira (adalimumab), used in the treatment of several autoimmune diseases, like Rheumatoid Arthritis (RA) or Crohn's disease (Mease 2007), Rituxan (rituximab) used in the treatment of leukaemias and lymphomas (Maloney et al. 1997) and Herceptin (trastuzumab) used in the treatment of breast cancer (Hudis 2007).

1.4 Antibody structure

Most vertebrate antibodies are Y-shaped molecules, composed of two chains – heavy and light. In this section, we focus on antibody structure and discuss which factors drive the binding properties of most known antibodies. In section 1.3.3 we stated that there are five antibody isotypes – IgG, IgA, IgD, IgE and IgM. Throughout this section, we discuss antibody structures from the point of view of the most common isotype – IgG. The other isotypes differ in the structure of the heavy constant region (see Figure 1.7).

1.4.1 Structural elements of an antibody

A human antibody is a Y-shaped dimer structure composed of four chains, two heavy and two light (see Figure 1.7). The two heavy chain copies are bound together through

several inter-chain disulphide bridges. Each heavy chain is composed of four domains – one Variable Heavy domain (VH) and three Constant Heavy domains (CH_1, CH_2 and CH_3). The light chain is composed of two domains – Variable Light (VL) and Constant Light (CL). The structure of an antibody is further divided into three fragments – the Fragment antigen-binding (Fab), Fragment variable (Fv) and Fragment crystallisable (Fc). The Fab segment can be detached from the rest of the structure, without unfolding, using an enzyme called papain (derived from papaya). The detached Fab preserves antigen-binding properties of an antibody (Berg et al. 2012). The binding is conveyed through the smaller Fv fragment. The Fc fragment, named for its propensity to readily crystallize, does not take part in the binding, but is responsible for a number of effector functions (Murphy et al. 2012), some of which were discussed in Section 1.3.2.

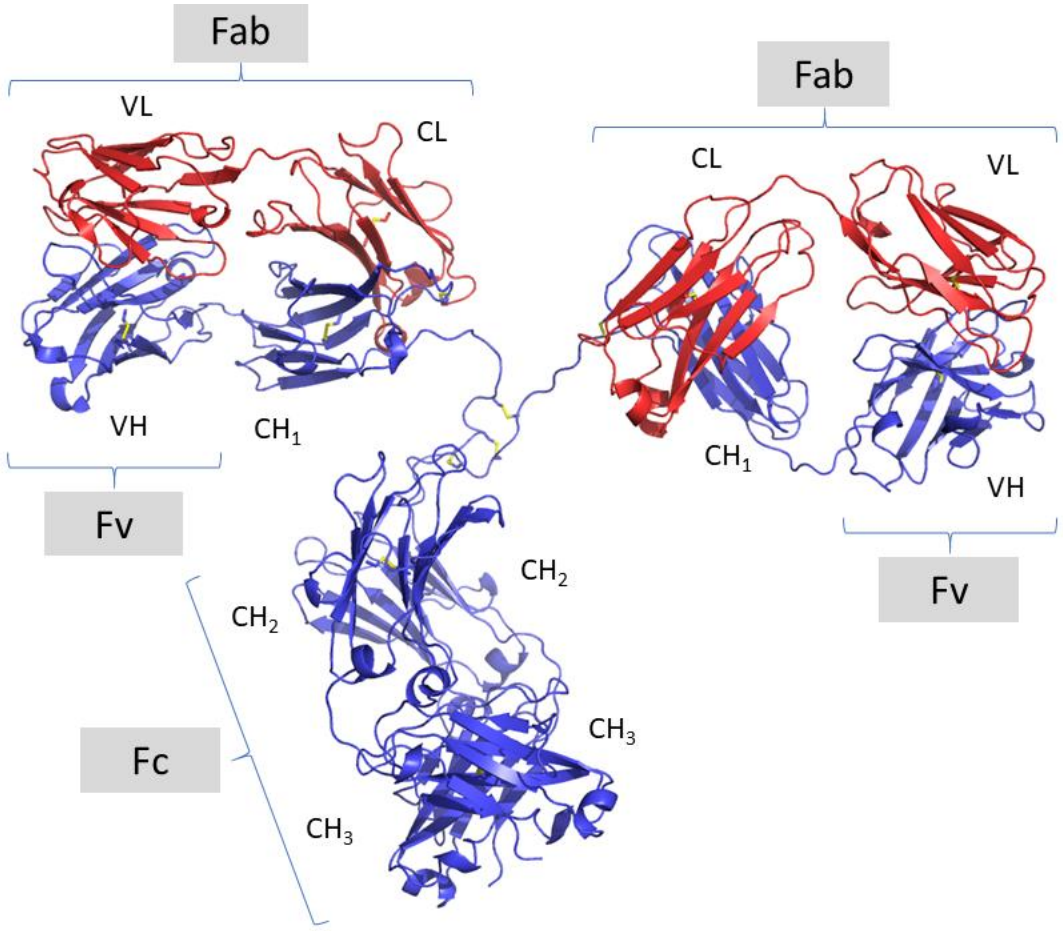

Figure 1.7 IgG structure. The schematic shows the crystallographic model of an entire antibody of IgG isotype (PDB id 1IGT). The light chain is shown in red and the heavy chain is shown in blue. The antibody fragments are labelled in grey. These are: Fragment antigen-binding (Fab), Fragment variable (Fv) and Fragment crystallisable (Fc). The binding properties of a typical antibody are determined by the Fv located at the ends of each arm of the Y-shaped structure. The domains of both chains are labelled. These are Variable Light domain (VL), Variable Heavy domain (VH), Constant Light domain (CL) and three types of Constant Heavy domain (CH_1, CH_2 and CH_3). A typical antibody is a dimer – the two arms of the antibody have the same sequence (there are exceptions to this rule – see Section 1.4.5).

1.4.2 Immunoglobulin fold

The six antibody domains (VH, VL, CL, CH_{1-3}, see Figure 1.7) share a common structural pattern, known as the immunoglobulin fold, prevalent among proteins participating in the

immune response (e.g. TCRs (Barclay 1999)). The immunoglobulin domain pattern consists of two antiparallel β-sheets (see Section 1.2.2), connected by disulphide bridges. The double β-sheet assembly is known as a β-sandwich. In each β-sheet the strands are arranged in a Greek key motif (see Figure 1.8).

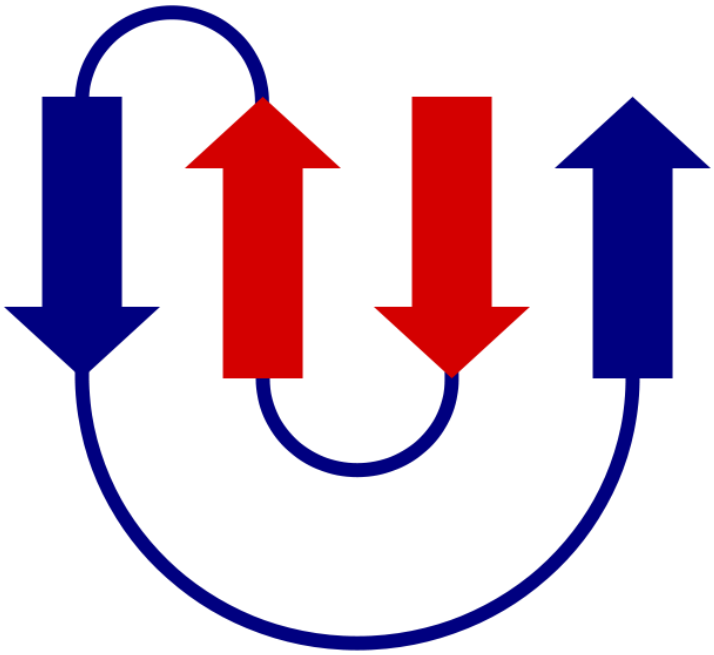

Figure 1.8 The Greek key topology. The schematic shows a cartoon representation of a Greek key motif consisting of four adjacent β-strands. The three leftmost strands are connected by short loops, known as β-hairpins, while the fourth strand is connected to the first via a longer loop. The image was created by Natelewis and was reproduced under CC BY-SA 3.0 licence (https://creativecommons.org/licenses/by-sa/3.0/).

In the Fv segment, the outer-facing loops, which connect the beta strands, contain the binding site of an antibody and determine its complementarity. They are known as the Complementarity Determining Regions (CDRs).

1.4.3 Complementarity Determining Regions

The binding properties of an antibody are primarily determined by the sequence and structure of just six loops called complementarity-determining regions (CDRs) (see Figure 1.9). Three CDRs are found on the light chain (CDR-L1 – CDR-L3) and three on the heavy chain (CDR-H1 – CDR-H3). The boundaries of CDRs were first defined in 1970 (Wu et al. 1970). Since then, a number of other definition have been developed, as part of the antibody sequence numbering schemes (see Section 1.5.1). Due to the importance of the CDRs, substantial efforts have been made to characterize them. The CDRs can be subdivided into the canonical CDRs (CDR-L1, CDR-L2, CDR-L3, CDR-H1, CDR-H2) and CDR-H3.

1.4.3.1 Canonical Complementarity Determining Regions

Comparison of antibody structures showed that the non-H3 CDRs (CDR-L1, CDR-L2, CDR-L3, CDR-H1, CDR-H2) form only a relatively small number of shapes, referred to as canonical classes (Chothia et al. 1987). A canonical class describes a set of loops that assume similar conformations, with the conformation being determined by the number and identity of the residues that constitute the loop and some residues in the framework region, adjacent to the loop. The theory of canonical classes postulates that the class of a loop can be identified by the presence of a few "key" residues at particular positions (Chothia et al. 1987). Thus, using canonical classes, it should be possible to predict the structure of a novel CDR, by classifying it using key features in its sequence.

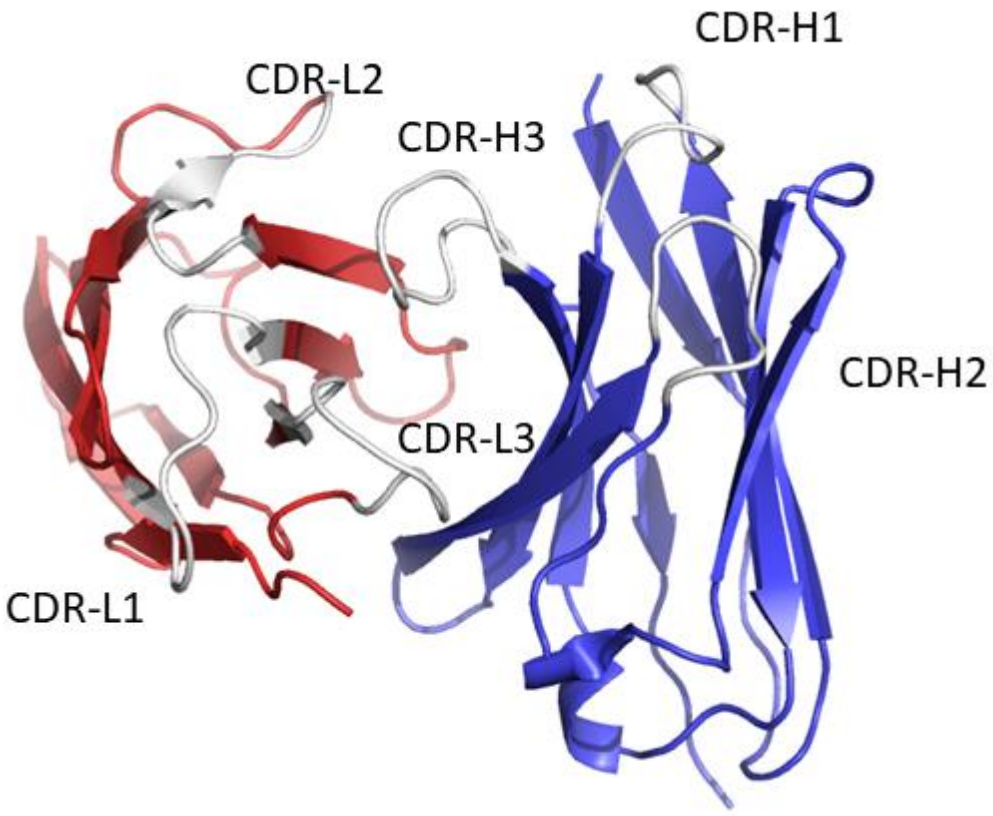

Figure 1.9 The Complementarity Determining Regions. The image shows the Fv fragment of an antibody with PDB id 1IGT with the six-complementarity determining region (CDR-L1, CDR-L2, CDR-L3, CDR-H1, CDR-H2, CDR-H3) highlighted in white. The CDRs have been selected according to the Chothia definition (Al-Lazikani et al. 1997) (see Section 1.5.1). Framework region of light chain is shown in red and framework of heavy chain is shown in blue. The CDRs determine the binding properties of an antibody and vary the most in sequence and structure. The CDR-H3 is the most variable of all CDRs, and typically forms the highest number of contacts with the antigen (Sela-Culang et al. 2013).

Since the original canonical class study of Chothia and Lesk (Chothia et al. 1987), the clustering of non-H3 CDRs into canonical forms has been extended several times (e.g. Chothia et al. 1989; Barre et al. 1994; Rees et al. 1994; Tomlinson et al. 1995; Martin et al. 1996; Al-Lazikani et al. 1997; Morea et al. 1997, 2000; North et al. 2011; Dunbar et al. 2014; Nikoloudis et al. 2014; Nowak et al. 2016).

The earliest clustering of CDR structures by Chothia and Lesk (Chothia et al. 1987) was performed with only five antibody structures and the comparison was done manually. In contrast, Martin and Thornton (Martin et al. 1996) created a fully automatic method for classification of CDRs into canonical forms, first clustering the structures in torsional space and then merging the clusters using root-mean square deviation (RMSD). Martin and Thornton (Martin et al. 1996) were also the first to note the limitations of the canonical model, in particular that sequence is not a perfect determinant of cluster membership. In the more recent study of North et al. (2011). CDR structures were clustered in torsional space, using the affinity propagation algorithm. This clustering is available as an online database (http://dunbrack2.fccc.edu/PyIgClassify/) (Adolf-Bryfogle et al. 2015).

There have also been studies of canonical shapes that involved only a subset of available structures. Some analyzed only specific chains (Chothia et al. 1992; Tomlinson et al. 1995; Chailyan et al. 2011) while others focused on individual non-H3 CDRs, in particular the CDR-L3 (Kuroda et al. 2009; Teplyakov and Gilliland 2014).

1.4.3.2 Heavy chain third Complementarity Determining Region (CDR-H3)

The CDR-H3 loop is distinct from the other five CDRs in that it has not been observed to form structural clusters (North et al. 2011). CDR-H3 is also more diverse in structure and sequence in comparison to the other CDRs (Weitzner et al. 2015). This is because the DNA coding region for this loop is at the interface between the V, D and J genes, which makes it more susceptible to modifications (see Section 1.3.4). Upon antigen binding, CDR-H3 typically becomes part of the core of the complex and forms the highest number of contacts with the antigen, out of all CDR loops (Sela-Culang et al. 2013).

Due to the importance of CDR-H3 there has been significant effort to understand its structural patterns (e.g. Reczko et al. 1995; Shirai et al. 1996, 1999; Morea et al. 1998; Oliva et al. 1998; Furukawa et al. 2001; Kuroda et al. 2008; Weitzner et al. 2015; Marks et al. 2017). The structure of the CDR-H3 loop can be separated into the "stem" region, containing residues close to the antibody framework and the "torso" region, containing residues away from the framework. The stem region consists mainly of residues coded for by the V germline and the J germline and is less structurally variable than the torso region, making it easier to structurally characterize (Teplyakov et al. 2016). In their work on CDR clustering, North et al. (2011). classified the stem region of CDR-H3, defined as the first three residues and the last four residues of the loop, into clusters.

1.4.4 Antibody framework and the VH-VL orientation

The non-CDR part of an antibody chain is known as the framework region. The framework region does not usually form many contacts with the antigen and is instead responsible for maintaining conformation of the CDRs. It is also less structurally variable than the CDRs. Nevertheless, changes to the sequence of a framework region can have both direct and indirect impacts on the affinity of the antibody; for example, affecting the structure and flexibility of the CDRs (Sela-Culang et al. 2013).

The contacts at the VH-VL interface influence the orientations between the two chains and, therefore, also have an indirect effect on the affinity of the antibody to its target (Dunbar et al. 2013). Due to the importance of the VH-VL orientation for antibody-antigen binding there has been significant effort to define a robust and consistent measure quantifying the orientation. Such measure can be relative, quantifying the difference in orientation between two Fv structures or absolute, describing the orientation without referencing other structures.

A commonly used absolute measure of VH-VL orientation is the packing angle described by Abhinandan and Martin (2010). The authors identified conserved beta-sheet strands in both heavy and light chain structures and defined their packing angle as the torsion between four points lying on those conserved beta sheets. Using the antibody structures available at the time they found that their packing angle spans values between -31.0° (for PDB id 1fl3) and -60.8° (for PDB id 1bgx) and identified 13 interface residues that are useful for predicting the angle (Abhinandan and Martin (2010)).

A different approach was developed by Dunbar et al. (2013). The authors described the VH-VL orientation by measuring the orientation between two planes, one fit through the conserved residues in the VH chain and one through the conserved residues in the VL chain. They quantified the orientation through five torsion angles (called HL, HC1, HC2, LC1, LC2) and a distance (dc). By analysing antibody crystal structures, the authors identified sets of interface residues that were predictive for each of the six orientation metrics Dunbar et al. (2013).

A commonly-used relative measure is known as orientation RMSD (Narayanan et al. 2009). Calculations of the orientation RMSD require two Fv structures as input. The score is then calculated by comparing two structural alignments. The first alignment is between the Cα atoms of the heavy chains and the second alignment is between the Cα atoms of the light chains. The orientation RMSD is calculated as the difference in light chain atom coordinates between the light chain position in alignment one and its position in alignment two (see Figure 1.10). The orientation RMSD has been used in many previous studies of VH-VL orientation (e.g. Colman et al. 1987; Li et al. 2000; Sela-Culang, Alon, and Ofran 2012). In Chapter 3, we use the orientation RMSD for quantifying differences in orientation between Fv regions of different antibodies.

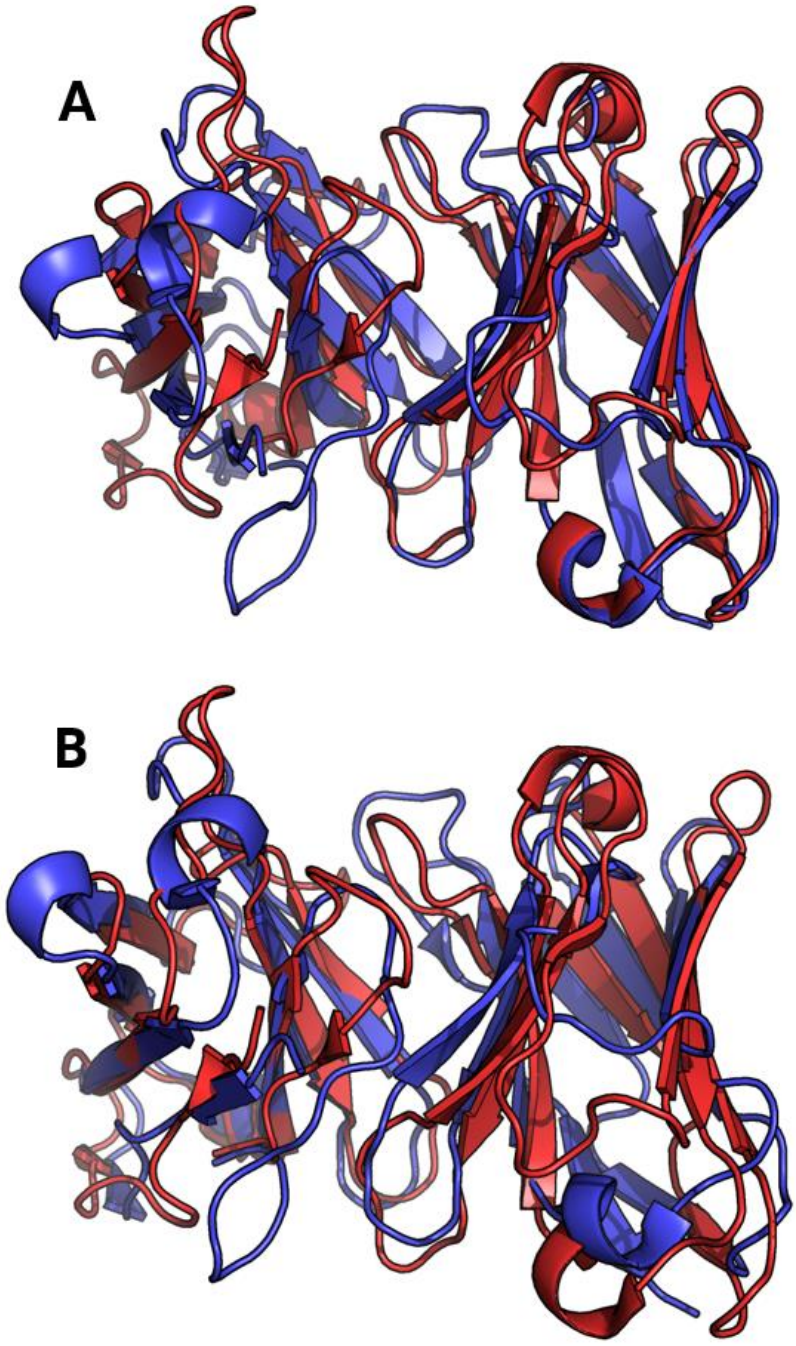

Figure 1.10 The calculation of orientation RMSD. The image shows two Fv regions, one from antibody with PDB id 1u6b (red) and the other from antibody with PDB id 2gfb (blue). The light chains are on the right while the heavy chains are on the left. First, the light chains are aligned (Panel A). Next, the heavy chains are aligned (Panel B). The orientation RMSD is calculated as the difference in light chain backbone atom coordinates between the light chain position shown in Panel A and the position in Panel B. In this example, the orientation RMSD in 9 Å.

1.4.5 Non-typical and engineered antibody structures

Throughout this section, we described antibody structure from the point of view of a typical human IgG antibody. It is important to note that some species are capable of producing antibodies containing structural elements very different from the ones discussed here.

One such example of an "exotic" structure type is found in bovine antibodies which often contain a very long CDR-H3, usually 50 to 60 residues long (F. Wang et al. 2013), composed of a "stalk" region and a "knob" region. The specificity of the bovine antibody is mostly conveyed through the knob region of the CDR-H3, with other CDRs not taking part in binding.

In Section 1.4.1 we stated that typical antibodies are dimers and that the two arms of the antibody usually share the same sequence. A notable exception from this rule is the IgG4 antibody which has been observed to exchange the antibody arms between different molecules (Aalberse et al. 2002). This continuous exchange causes the IgG4 to behave more like a monomer (Aalberse et al. 2002).

Another example is a single-chain camelid antibody whose Fab region consists of a single variable heavy chain domain. A camelid heavy chain is soluble, meaning it can exist in solution without being attached to a light chain. Such single-domain antibodies are very stable - some remain folded at temperatures exceeding 80°C (Goldman et al. 2017). Due to this high stability and small size, artificially developed single-domain antibodies (sdAbs or Nanobodies) have attracted the interest of therapeutic researchers (Hussack et al. 2011; Van de Broek et al. 2011).

Non-standard antibody structures can be artificially engineered for therapeutic purposes. Two examples of this are bispecific antibodies and single-chain variable

fragments (scFvs). Bispecific antibodies consist of two distinct Fab fragments connected together, with each Fab fragment being specific for a different antigen (Chames et al. 2009). One application of such bispecific antibody is to bring a cytotoxic killer T-cell near a tumour cell (Chames et al. 2009). An scFv consists of a variable region of a heavy chain and a variable of a light chain fused together through a peptide linker (Huston et al. 1988). ScFvs have been successfully used in phage display, by displaying the fragment on the surface of a virus (McCafferty et al. 1990).

1.5 Computational characterisation of antibody structure

In the previous section, we discussed various elements of antibody structure. Here, we discuss computational methods for predicting three-dimensional structure of an antibody Fv region. These methods use concepts introduced in Section 1.2.4. The accuracy of antibody structure prediction methods is periodically evaluated by Antibody Modelling Assessment (AMA) (Teplyakov, Luo, et al. 2014).

1.5.1 Antibody sequence numbering

While not strictly a structure prediction method, antibody numbering serves to annotate residues from antibody sequence with their approximate structural position and function, without knowledge of the full antibody structure. For example, a heavy chain residue with number 30, according to Chothia numbering, would always be part of CDR-H1 (without conveying any information about CDR-H1s structure), while a heavy chain residue number 60 would always be part of the framework etc. Many antibody numbering schemes have been created (e.g. Kabat et al. 1983; Chothia et al. 1987, 1989; Al-Lazikani et al. 1997; Honegger et al. 2001; Lefranc et al. 2003; Abhinandan et al. 2008; North et al. 2011). One of the earliest antibody numbering schemes was introduced by Kabat et al. (1983). Kabat numbering was created by aligning known

antibody sequences, without considering the structure (Kabat et al. 1983). To include structural consideration, a new numbering scheme was introduced by Chothia et al. in 1987 (Chothia et al. 1987). The Chothia numbering scheme was updated multiple times to better annotate structurally equivalent positions (Chothia et al. 1989; Al-Lazikani et al. 1997). In both the Chothia scheme and the Kabat scheme heavy and light chains are numbered separately, according to different rules. In 2003 a unified numbering scheme was released, known as the IMGT scheme, which was also extended to other variable Immunoglobulin domains (Lefranc et al. 2003).

Antibody sequences can be automatically numbered using program ANARCI (Dunbar et al. 2015). ANARCI numbers a query sequence by aligning it to an HMM model constructed from a database of germline (see Section 1.3.4.1) sequences. The positions of the HMM alignment are annotated with numbers from Kabat (Kabat et al. 1983), Chothia (Al-Lazikani et al. 1997), Extended Chothia (Abhinandan et al. 2008), IMGT (Lefranc et al. 2003) and AHo (Honegger et al. 2001) schemes. ANARCI also annotates chain type, species, V-germline and J-germline for the query sequence (Dunbar et al. 2015).

Throughout this thesis, we use ANARCI software to number antibody sequences according to Chothia (Al-Lazikani et al. 1997) and IMGT schemes (Lefranc et al. 2003).

1.5.2 Framework region and VH-VL orientation

The framework region of an antibody consists of an Fv fragment without the CDR loops. The framework regions show a high degree of structural similarity and can usually be accurately modelled using homology methods (Weitzner et al. 2017). It has been shown that when a template with sequence identity above 0.80 is available, there is 75%

chance of modelling the heavy chain with RMSD better than 1Å and 88% chance of modelling a light chain with RMSD better than 1Å (Leem et al. 2016).

Modelling the orientation between the heavy and light chain is much more challenging (Dunbar et al. 2013) and several methods have been developed specifically for this purpose. The simplest method for modelling the VH-VL orientation involves choosing a template with the highest sequence identity over the whole framework (e.g. Fasnacht et al., 2014). This method is often sufficient, if a template with high sequence identity is available, however it ignores the fact that most residues in the Fv structure do not contribute to the VH-VL interaction.

A different approach involves using machine learning methods to predict the orientation (e.g. Abhinandan and Martin 2010; Bujotzek et al. 2015; Dunbar et al. 2013). The advantages of these methods include automatic identification of interface residues important for orientation prediction, and their ability to make a prediction in every case, irrespective of template availability. The main disadvantage is a low volume of training data - as of August 2017, at 99% sequence identity, there are only ~800 Fv structures available in the PDB database.

Recently, a different method for predicting the orientation was created by Marze, Lyskov, and Gray (2016). The authors employed the protein-protein docking routines within the Rosetta software package (Chaudhury et al. 2011) to create multiple initial orientation poses for antibody modelling. These poses were then used as input for CDR modelling and scored and selected only after the CDR modelling step. The authors asserted that their method could construct accurate models for 33 out of 46 benchmark structures, which is an impressive performance for an *ab initio* method (Marze et al. 2016). Nevertheless, the accuracy comes at a significant cost in computing time

(~1,440 CPU hours per target) and is not suitable for handling large volumes of antibody sequence data.

In Chapter 3 we create a novel, high-throughput methodology for predicting VH-VL orientation using sequence identity calculations over the VH-VL interface residues.

1.5.3 The Complementarity Determining Regions

Five out of six CDRs (CDR-L1, CDR-L2, CDR-L3, CDR-H1, CDR-H2) form only a small number of conformations, known as canonical classes. The canonical classes can be identified through clustering available CDR structures (see Figure 1.11) and new loops can be classified to the canonical classes by identifying key positions in the loop sequence (Chothia et al. 1987). This classification can be done by manual inspection (Chothia et al. 1989) or using a machine-learning algorithm, such as Hidden Markov Model (HMM) (North et al. 2011; Nowak et al. 2016). In Chapter 2 we create a novel clustering of CDR structures and describe a machine learning method to match structurally uncharacterised loops into clusters.

The structure of a CDR can also be predicted using homology modelling (e.g. Hildebrand et al. 2009; Choi et al. 2010; Holtby et al. 2013; Messih et al. 2015). In Chapter 5, we describe how the homology modelling FREAD software (Choi et al. 2010) can be used to assign structural templates to CDR sequences.

As discussed in Section 1.4.3.2 the structures of CDR-H3 loops do not form canonical classes. Because of this, more sophisticated loop structure prediction methods need to be used to model the CDR-H3 loop. In cases where structural templates are available, homology modelling of CDR-H3 can yield accurate predictions (Leem et al. 2016). When templates are not available, the loop structure must be predicted using *ab initio*

algorithms (e.g. Xiang et al. 2002; Soto et al. 2008; Stein et al. 2013; Liang et al. 2014). Ab initio algorithms use heuristic methods to constrain the conformational search space from which structures are sampled. They are usually much slower than template-based methods – a template based method can often return a prediction in a fraction of a second, while a *de novo* method might take hundreds of CPU hours to accomplish the same task. In Chapter 3 we describe a hybrid CDR-H3 modelling software called Sphinx which combines concepts from homology and *de novo* structure prediction methods.

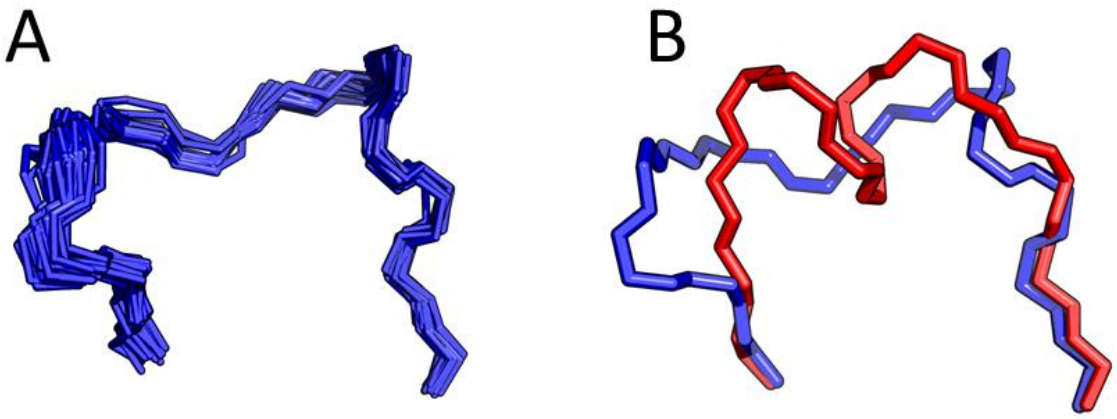

Figure 1.11 CDR canonical classes. The figure illustrates the structural homology of loops in a canonical class. Panel A shows alignment of 7 sequence unique CDR-L1 loops of length 14, falling into the same canonical class (L1-14-A class, see Chapter 2). The aligned CDRs are shown in blue. Panel B shows a structural alignment of two CDR-L1 loops of length 14 falling into different canonical classes. A loop from L1-14-A class is shown in blue, while a loop from L1-13,14-A class is shown in red. The structural similarity of CDRs in a canonical class makes it possible to use the concept for modelling novel CDR sequences.

1.5.4 Antibody-antigen docking

A structural model of an antibody's Fv region can be an invaluable source of information. Nevertheless, the functionality of an antibody is conveyed through its interactions with

the antigen. Therefore, to better understand the activity of an antibody, the three-dimensional antibody-antigen complex needs to modelled.

The field of predicting the structure of protein-protein complexes is known as protein docking. Since the inception of the field, a plethora of docking methods have been created (e.g. Katchalski-Katzir et al. 1992; Schneidman-Duhovny et al. 2005; Chaudhury et al. 2011). These docking algorithms usually take as input two solved structures of unbound proteins and attempt to predict structure of the complex. If structures of the query proteins are not available, docking can be performed with structural models, but the results tend to be less accurate (Tovchigrechko et al. 2002). Like protein structure prediction methods, docking can be performed using homology or *ab initio* approaches. The homology methods (see Section 1.2.4) search for templates of structurally characterised complexes, similar to the complex of interest. The *ab initio* methods (see Section 1.2.4) sample structural models of the complex, known as docking poses, from a conformational space. In the specific case of antibody-antigen modelling, the complex must be modelled using *ab initio* methods, since the binding modes of antibodies are usually not related to each other or to other protein-protein complexes (Krawczyk et al. 2013). Nevertheless, it has been shown that over 90% of the antibody residues observed to contact the antigen are part of the CDRs and the neighbouring framework residues (Kunik, Ashkenazi, et al. 2012; Kunik, Peters, et al. 2012). Restricting the antibody binding sites, sampled by the docking algorithm, to these residues confines the conformational search space, which improves the efficiency of the docking procedure (Sircar et al. 2010).

The sampling of docking poses can be performed exhaustively or randomly. The commonly-used approach for performing exhaustive search involves representing

different orientations of the binding partners' structures as points on a grid and scoring the relative orientations using convolution functions (Katchalski-Katzir et al. 1992). The convolutions can be efficiently calculated using Fast Fourier Transform (FFT), which decreases the complexity of the algorithm (Katchalski-Katzir et al. 1992), making the process computationally tractable. The disadvantage of the exhaustive search algorithms is that the grid representation of proteins inherently coarse-grains the structural features and does not allow for any flexibility in side chain or main chain conformation (Katchalski-Katzir et al. 1992). The latter point is especially important, since it has been shown that protein binders often exhibit induced fit behaviour (Weikl et al. 2009). The random search methods typically utilize a Monte Carlo trajectory, where random perturbations to the previous pose are accepted or rejected based on predicted energy and Metropolis criterion (Chaudhury et al. 2011). The advantages of random search algorithms are that they can incorporate side chain and backbone flexibility, and that the energy calculations do not have to be represented through convolutions (Gray et al. 2003). One of the disadvantages is that the search might converge to a suboptimal solution. In addition, the running time is usually longer than for exhaustive search methods. The accuracy of the contemporary docking methods is periodically assessed in the CAPRI (Critical Assessment of PRedicted Interactions) competition (Méndez et al. 2003).

In Chapter 5, we use two docking algorithms – the ZDOCK algorithm (Pierce et al. 2011) and the docking method implemented in Rosetta (Chaudhury et al. 2011) – to generate antibody-antigen poses as part of a computational antibody design pipeline.

1.5.5 Sources of antibody data

The antibody structure prediction methods described above could not have been designed and benchmarked without the publicly available antibody sequence and structure data. Here, we describe some openly accessible antibody databases.

1.5.5.1 Structural data

Atomic coordinates of structurally characterised proteins are deposited in the PDB database (Berman et al. 2000). The PDB contains a number of antibody structures, but extracting this data can be laborious. In addition, the antibody structures obtained from the PDB are only annotated in a generic manner, without any antibody-specific information. To make antibody structures more accessible, in 2013 Dunbar et al. designed the Structural Antibody Database (SAbDab) (Dunbar et al. 2014). SAbDab regularly parses newly added PDB files, searching for antibody sequences. The matched antibody-containing files are annotated with such information as species of origin, germline genes, chain identifiers, antibody-antigen pairings, affinity etc. The files are also numbered (see Section 1.5.1) and annotated with structural features (VH-VL orientation, CDR canonical classes etc.). Throughout this thesis, we make extensive use of antibody structural data obtained from SAbDab database.

1.5.5.2 Sequence data

In recent years, there has been an exponential growth in the number of available antibody sequences, caused mainly by the development of cheap and high-throughput Next-Generation Sequencing (NGS) techniques. The state-of-the-art methods allow sequencing of most of the 10^6 B-cells contained within the commonly-used 10-ml blood draw (DeKosky et al. 2014). The increased availability of high-volume antibody sequence data has led to creation of several public antibody sequence databases, such

as the Kabat database (Johnson et al. 2001), IMGT/LIGM-DB (Giudicelli et al. 2006), abYsis (Martin 2001), VBASE2 (Retter et al. 2005) and the DIGIT database (Chailyan et al. 2012). These databases contain on the order of 10^5 – 10^6 antibody sequences. Given that, at the time of writing this text, there are only ~2,000 crystal structures of antibodies available (Dunbar et al. 2014) the structural characterisation of such sequence data could provide invaluable insights into the available antibody space. In Chapter 2 of this thesis we predict CDR canonical classes for large volume of CDR-L3 sequence data. In Chapter 4, we use a large dataset of IgM sequences to predict existence of novel canonical classes. In Chapter 5, we use the same dataset to construct a novel pipeline for computational antibody design.

1.6 Computational antibody design

There are a number of methods for developing antibodies experimentally. Nevertheless, these methods are often time-consuming and expensive, making the concept of computational therapeutic development attractive. Despite the huge leaps in experimental antibody design (see Section 1.3.6), the field of computational antibody design remains relatively nascent, due to enormous complexity and high error rate of the steps involved in the design process. Recent decades have seen a continued exponential growth in available computational power (Denning et al. 2016) and significant increases in the accuracy of antibody structure prediction methods (Leem et al. 2016; Marks et al. 2017; Weitzner et al. 2017). These trends make the prospect of developing a computational pipeline for antibody design more realistic.

Most computational antibody design studies to date focused on re-designing antibody-antigen binding complexes using a crystal structure of a complex as a starting point (e.g. Clark et al. 2006, 2009; Lippow et al. 2007; Tharakaraman et al. 2013; Kiyoshi et al.

2014; Xu et al. 2014). Such re-design protocols involve utilization of antibody-antigen affinity prediction methods to identify beneficial structural changes and amino acids substitutions (Kuroda et al. 2012).

In recent years a number of programs (e.g. Pantazes et al. 2010; Li et al. 2014; Lapidoth et al. 2015) have been developed for the purpose of *de novo* antibody design.

The quality of the computational antibody design algorithms is typically assessed by comparing the computationally designed binders against structurally characterised antibody-antigen complexes.

One comparison method involves aligning the designed complex with the structurally characterised one and measuring the RMSD between interface atoms of the antigen in the designed position and in the experimentally identified position (e.g. Li et al. 2014). This measure is known as the interface RMSD (I_RMSD) and is also used to assess the quality of docking poses in the CAPRI competition (Méndez et al. 2003). Typically, docking poses with I_RMSD < 1.0 Å are classified as correct, poses with 1.0 Å < I_RMSD < 2.0 Å are classified as medium, poses with 2.0 Å < I_RMSD < 4.0 Å are classified as acceptable and poses with I_RMSD > 4.0 Å are classified as incorrect.

A different comparison method involves calculating computational metrics that describe various properties of the designed complex (e.g. predicted affinity of binding, solvent accessible surface area or surface complementarity) and comparing them to equivalent metrics obtained for real antibody-antigen complexes (e.g. Lippow et al. 2007; Lapidoth et al. 2015). This class of quality measures is based on the assumption that the structurally characterised antibody-antigen complexes have desirable properties for optimal binding which should be built into the computational designs.

The choice of quality assessment method depends on the application of the computational design method. If the goal of the study is to re-design an existing antibody-antigen complex using single amino acid substitutions (e.g. Clark et al. 2006, 2009; Lippow et al. 2007), then energetic considerations typically take priority over structural ones, as it can be assumed that the correct epitope and binding pose are known *a priori* and are unlikely to change during the re-design process (Clark et al. 2006). In these studies, the optimization objective usually takes form of an approximation to the antibody-antigen binding affinity, such as calculated by the CHARMM potential (MacKerell et al. 1998) or Rosetta scoring function (Chaudhury et al. 2011). Conversely, when the epitope or binding pose cannot be assumed to be known *a priori,* the structural comparison methods, such as the I_RMSD, are typically used to show that the computational design algorithm can correctly delineate the epitope (e.g. Pantazes et al. 2010; Li et al. 2014; Lapidoth et al. 2015).

In the following sections, we describe in more detail some of the computational antibody design algorithms, available at the time of writing this thesis.

1.6.1 Antibody re-design

Antibody re-design methods are based on the idea that it is possible to modify an antibody's sequence and structure without loss of stability. The methods involve the use of energy functions to identify binding site modifications which increase the predicted affinity of an input antibody to its antigen

One of the earliest re-design studies (Lippow et al. 2007) focused on re-designing complexes formed between three antibodies and their antigens. These antibodies were: anti-epidermal growth factor receptor drug cetuximab (Sato et al. 1983), anti-lysozyme antibody D44.1 and anti-lysozyme antibody D1.3 (Lippow et al. 2007). The

authors mutated all the CDR residues (as defined by Kabat numbering) on each of their antibody targets to all 20 amino acids, except Proline and Cysteine, estimating the change in binding energy using the CHARMM PARAM22 potential (MacKerell et al. 1998). The effects of top scoring mutations were re-estimated using more computationally expensive Poisson-Boltzmann continuum electrostatics (Sharp et al. 1990). Then, the best scoring point mutations for each target were experimentally characterized. Finally, the point mutations shown to increase binding affinity were combined to produce an improved binder. Using their methodology, Lippow et al. improved the binding affinity of D44.1 tenfold and the affinity of cetuximab by 140-fold (Lippow et al. 2007). The antibody D1.3 gave few opportunities for electrostatic improvement and the authors concluded it might be an anomalous antibody (Lippow et al. 2007).

In 2009, the binding site of an anti-VLA1 antibody (PDB id 1MHP (Karpusas et al. 2003) was re-designed by replacing the original CDR-L1 loop with CDR-L1s from other solved antibody structures (Clark et al. 2009). The residues of the replacement CDR-L1s were subjected to mutations, using DEZYMER software (Looger et al. 2003), in search of low-energy sequences. The best-scoring mutants were expressed in *E. coli* and the affinity to the antigen was measured. All mutants successfully folded but the affinity to the target decreased significantly, with the affinity of the best mutant decreasing 100 times in comparison to the wild-type antibody. To explain their findings the authors crystallized the structure of the highest-affinity mutant and discovered that the transformed antibody forms dimers by domain swapping and that the structure of CDR-H3 became delocalized. Nevertheless, the transplanted CDR-L1 maintained the expected structure on the new framework.

The impact of growth in available computational power was illustrated in 2014 (Kiyoshi et al. 2014). The starting point in this study was a crystal structure of a complex between antibody 11K2 and its antigen, the anti-monocyte chemoattractant protein 1 (MCP-1) (Reid et al. 2006). Each CDR position was subject to mutations to all amino acids (1,178 mutations in total). For every such mutation the authors generated 100 models using the MODELLER software (Eswar et al. 2006) resulting in 117,800 models (in comparison, a study carried out in 2005 (Clark et al. 2006) generated only 500 models). The energy of each complex was then estimated using the module Einteract of the MOE package (http://www.chemcomp.com) (Kiyoshi et al. 2014). In addition, for energy comparison purposes, 1000 models of the original complex were built using the same methodology. Twelve mutations (seven on the light chain and five on the heavy chain), expected to improve the binding affinity, were selected for experimental investigation (Kiyoshi et al. 2014). The selected mutants were expressed in *E. coli* and their affinities were measured. All five heavy chain mutations were false positives, but five out of the seven light chain mutants resulted in an improved affinity to the target.

1.6.2 *De novo* antibody design

Recently, a number of programs have been developed with the aim of computationally designing an antibody for a novel target (Pantazes et al. 2010; Li et al. 2014; Lapidoth et al. 2015).

OptCDR algorithm is one of the first *de novo* pipelines to be created (Pantazes et al. 2010) (see Figure 1.12). First, the method finds combinations of CDR structures that maximize predicted affinity to a selected antigen epitope. This step is followed by modifications of the CDR sequences and backbone refinement. The program does not produce full Fv structures; instead it generates libraries of CDR sequences which need to

be grafted onto frameworks to produce binders (Pantazes et al. 2010). The software was tested by designing antibodies for three targets: the hapten fluorescein, a peptide from the capsid of hepatitis C and the VEGF protein (Pantazes et al. 2010). The designed models were found to have computational metrics similar to those of the real antibodies, however no experimental validation was carried out (Pantazes et al. 2010).

The same group later released an updated version of OptCDR called OptMAVEn (Li et al. 2014) (see Figure 1.12). The new version of the software follows a similar protocol to OptCDR, but instead of testing combinations of CDR structures the program combines the Variable structure segment with the CDR-3 structure and with the Joining segment structure, separately for each chain (Li et al. 2014). OptMAVEn also includes a novel antigen docking protocol which, in 96% of test cases, could position the antigen within 4.0 Å to the binding pose observed in the crystal structure, as measured by the I_RMSD (Méndez et al. 2003). The structural assembly and docking stage is followed by simultaneous affinity maturation and humanization stage during which the sequence of the antibody is modified to increase strength of binding to the antigen and to resemble natural human sequences (Li et al. 2014). The program was tested by designing antibodies for two antigens: envelope glycoprotein gp120 (gp120) and hemagglutinin (HA). The properties of point mutations introduced by the algorithm were consistent with what is observed during *in vivo* affinity maturation (Li et al. 2014). The OptMAVEn was validated experimentally in 2017, where the authors designed antibody binders against a peptide antigen (Poosarla et al. 2017).

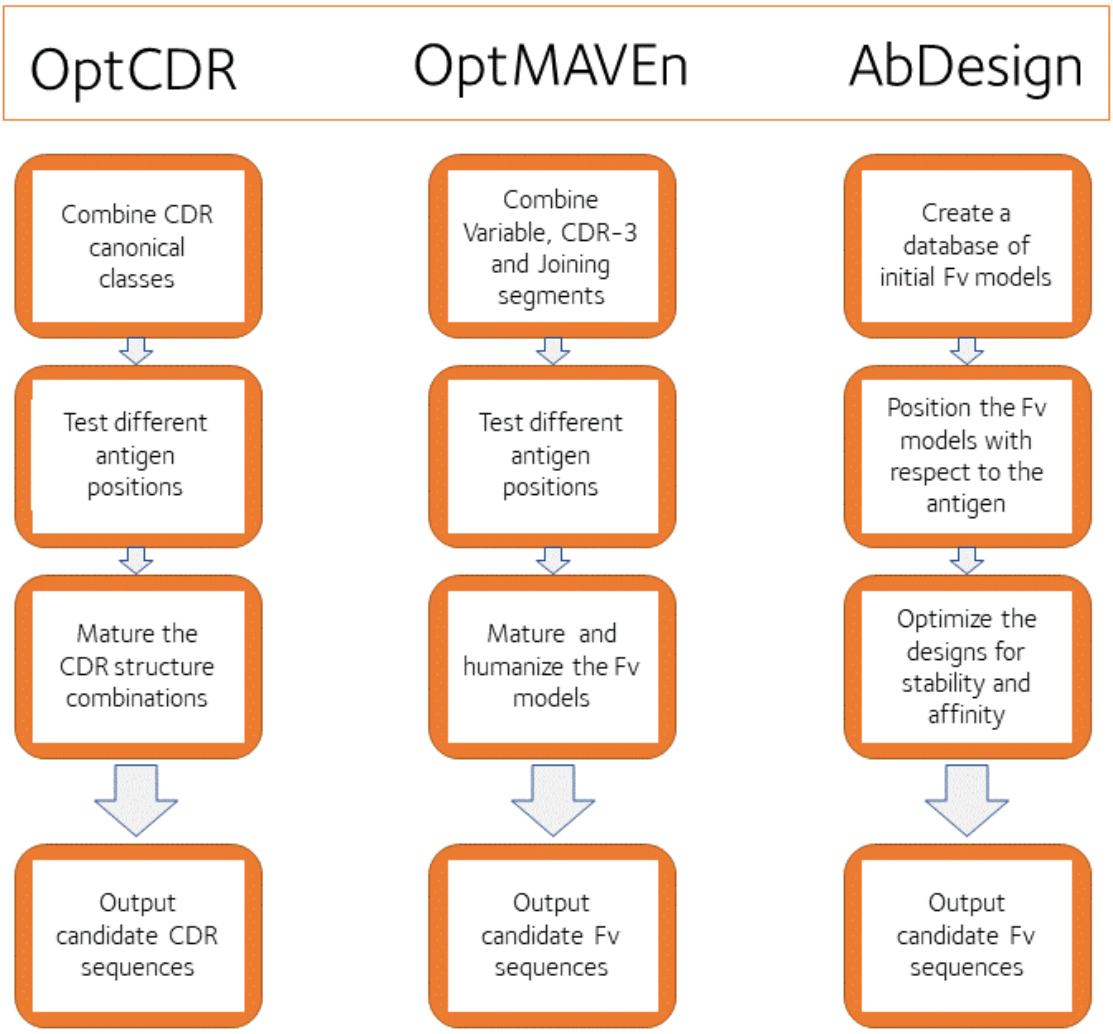

Figure 1.12 Computational *de novo* antibody design algorithms. The schematic shows the workflow of three computational antibody design algorithms: OptCDR (left, (Pantazes et al. 2010)), OptMAVEn (middle, (Li et al. 2014)) and AbDesign (right, (Lapidoth et al. 2015)).

In 2015, the AbDesign algorithm (Lapidoth et al. 2015) (see Figure 1.12) was presented. First, the method constructs a database of allowable torsion angles for each antibody segment (VH, VL, CDR-H3, CDR-L3) by clustering the available antibody structures (Lapidoth et al. 2015). The structural cluster centres are then used to create a library of 4,500 initial antibody conformations, by applying the torsion angles to a template antibody 4m5.3 (PDB id 1X9Q (Midelfort et al. 2004)). Alongside the torsion database, a corresponding Position Specific Scoring Matrix (PSSM) is built, specific for a given

structural cluster (Lapidoth et al. 2015). To design an antibody, AbDesign takes as input a crystal structure of an existing antibody-antigen complex (Lapidoth et al. 2015). First, the 4,500 initial conformations are aligned to the true antibody structure. The resulting poses are then randomized using a centroid mode of the Rosetta docking protocol (Chaudhury et al. 2011) and redesigned by perturbing the backbone structure and sequence in a simulated annealing trajectory, scoring the designs simultaneously on predicted binding energy and stability (Lapidoth et al. 2015). The resulting matured designs are then filtered based on several predicted metrics, such as shape complementarity and buried surface area (Lapidoth et al. 2015). AbDesign was validated by producing antibody designs for nine targets with known antibody-antigen complex structure (Lapidoth et al. 2015). In five cases the sequence identity of the designed paratope was above 30% and in four of those the designed binding site was within 1 Å of the natural antibody (Lapidoth et al. 2015).

These studies were the first attempts at *de novo* antibody design. They showed that it is possible to design antibodies that share common features with the real antibodies whose structures have been deposited in the PDB database. In Chapter 5, we develop a pipeline for *de novo* computational antibody design which uses a large NGS dataset of human IgM sequences as a starting point.

1.7 Outline of the thesis

This first chapter gives a theoretical background to the work conducted for this thesis. In this section, we briefly outline the contents of the other chapters.

1.7.1 Chapter 2

In Chapter 2 we describe the development of a novel, length-independent method for classifying antibody CDRs into canonical classes. First, we show that by clustering CDRs in a length-independent manner we can capture a broader range of structural similarities, than using a standard length-dependent approach. Next, we demonstrate that our length-independent canonical classes improve accuracy of classifying novel loop sequences into structural clusters. Then, we give a biological explanation for the existence of length-independent structural similarities, by discussing how they can arise through antibody diversity-generating processes. Finally, we demonstrate the power of our method by classifying CDR-L3 sequences from three large antibody sequence databases into clusters. Using our length-independent approach, we could assign canonical class to ~20% more sequences than using a standard length-dependent method (Nowak et al. 2016). The contents of this chapter were published as a research article (Nowak et al. 2016) and contributed to a separate article (Dunbar et al. 2016).

1.7.2 Chapter 3

In Chapter 3 we considered the problem of CDR-H3 structure prediction and VH-VL orientation prediction. First, we discuss improvements to decoy ranking protocol implemented in CDR-H3 structure prediction software Sphinx, developed by another member of our research group (Marks et al. 2017). First, we describe the development of a consensus machine-learning decoy ranking method and discuss why it does not improve the ranking in comparison to constituent ranking protocols. Next, we propose a structural relaxation method, which removed structural clashes from individual CDR-H3 decoys, improving the performance of decoy ranking algorithms. Finally, we show that crystallographic hydrogen bonds formed by target loops impact the accuracy of loop

structure prediction and that benchmark structures containing these contacts should be removed from the set. In the second part of Chapter 3, we describe the development of a high-throughput homology method for predicting VH-VL orientation. The algorithm assigns orientation through calculations of sequence identity between sets of VH-VL interface residues. By expressing the identity calculations as sparse matrix multiplication, we show that we can assign orientation to billions of potential Fv sequences in tractable time. The research conducted for this chapter contributed to two research articles (Marks et al. 2017; Parks et al. 2017).

1.7.3 Chapter 4

In Chapter 4 we describe the analysis of sequence patterns in a large dataset of antibody CDR sequences. First, we describe the development of a machine-learning autoencoder algorithm for grouping sequences based on statistical correlations between constituent residues. Next, we introduce an artificial dataset of CDR sequences and test our method at its ability to separate the pre-encoded sequence patterns. We found that our algorithm correctly identifies the encoded relations. Then, we show how we benchmarked our method by clustering CDR sequences from a large NGS dataset of antibody sequences mixed together with structurally characterised sequences analysed in Chapter 2. We compare our sequence clustering with canonical classes discovered in Chapter 2 and find that our algorithm correctly identifies structurally-relevant amino acid patterns. Finally, we describe sequence clusters without associated structural data and postulate that they might be representative of previously-unseen canonical classes. At the time of writing this thesis, the existence of these novel classes is being experimentally investigated by our industrial collaborators.

1.7.4 Chapter 5

In Chapter 5 we describe the development of a novel pipeline for computational antibody design. The method is based on using a set of ~15,000,000 sequences of human IgM chains as a cornerstone from which we derive a structurally diverse Antibody Model Library (AML) containing ~20,000 models. We describe the properties of the AML and show that the constituent models have similar structural features to the structurally characterised antibody therapeutics. Then, we describe how we panned this AML against four epitopes of Hen-egg lysozyme, producing ~100,000 candidate binding poses per epitope. Following that we describe how we used a custom machine-learning ranking method to select 100 most promising poses from this set of ~100,000 poses. We show that in each of the four cases, the set of 100 poses contained binding modes homologous to structurally characterised antibody-lysozyme complexes. Finally, we describe how we computationally matured a set of 100 poses designed for one of the lysozyme epitopes. We show that the matured designs contained structural features similar to real therapeutic binders. Finally, we describe how we predicted the propensity of our designed models to cause an immunogenic response and that by basing our design methodology on a set of human sequences, we were able to circumvent the immunogenicity problem. The set of anti-lysozyme designs is currently being experimentally tested by our collaborators at UCB Pharma with results pending.

1.7.5 Chapter 6

In the last chapter of the thesis, we summarize the presented results, deliberate on conclusions stemming from those results and discuss potential future directions.

2 LENGTH-INDEPENDENT STRUCTURAL SIMILARITIES ENRICH THE ANTIBODY CDR CANONICAL CLASS MODEL

2.1 Introduction

The contents of this chapter have been reproduced from a publication (Nowak et al. 2016).

As mentioned earlier, the binding properties of a typical antibody are determined by sequence and structure of six loops known as Complementarity Determining Regions (CDRs). Three CDRs are on the light chain (CDR-L1, CDR-L2, CDR-L3) and three are on the heavy chain (CDR-H1, CDR-H2, CDR-H3). It was shown that five out of six CDRs (CDR-L1, CDR-L2, CDR-L3, CDR-H1, CDR-H2) form only a limited number of conformations, known as canonical classes. In the earliest CDR canonical class study Chothia and Lesk (Chothia et al. 1987) noticed that there are CDR loops that, despite differences in length, are more structurally similar to each other than to other CDR loops

of the same length. The clustering method used by Martin and Thornton (Martin et al. 1996) allowed for comparison between loops of different length, but all the clusters discovered by the authors contained CDRs of only a single length. Most of the later clusterings were performed under the assumption that CDRs of different length are structurally distinct. Here, we quantify the structural similarities between loops of different lengths and create a methodology to find length-independent structural clusters of CDRs. We show that these length-independent clusters contain a larger number of unique sequences and are better able to predict structure from sequence than their length-dependent counterparts.

The latter result emphasizes the fact that the structural relationships between different length CDRs are based on sequence patterns. Using our length-independent structural clusters, we identified the most common causes of similarity between loop structures of different lengths. We demonstrate the impact of our study by analysing the cluster membership of CDR sequences from next-generation sequencing datasets. We show that by taking into account the structural similarities between loops of different length, we are able to classify significantly more CDR sequences into structural clusters.

2.2 Methods

2.2.1 Choice of CDR definition

For this study, we used the Chothia definition of CDR loops (Chothia et al. 1987) for all CDR types except for CDR-H2, where two residues before the N-terminus were also included. This choice was made as we tested if extending Chothia defined CDRs by up to three residues at either end would change the clustering results, especially the prediction accuracy (see methods Section 2.2.5). A change in length only made a statistically significant change to the results for CDR-H2, where it improved prediction accuracy. The

resulting boundaries of each CDR in Kabat-Chothia numbering are as follows: CDR-L1: 24-34, CDR-L2: 50-56, CDR-L3: 89-97, CDR-H1: 26-32, CDR-H2: 50-56, CDR-H3: 95-102.

2.2.2 Data selection

The data set was built from the 1833 antibody PDBs (www.rcsb.org) (Berman et al. 2000) available in the SAbDab database as of September 2014 (Dunbar et al. 2014). Antibody structures solved using methods other than X-ray crystallography and those solved with a resolution above 2.8 Å were removed from the dataset. Structures of CDR loops were extracted from the remaining PDBs along with their anchors, five residues before the N-terminus and five after the C-terminus. CDR structures were removed from the dataset if they had atoms missing from the loop or anchor region or if they contained backbone atoms with b-factors above 80Å^2 or equal to zero. Loops with identical sequences resulting from solving the structure of the same antibody multiple times were not removed because they can have different structures.

We use the following nomenclature for our structures: four letters for the PDB code of an antibody, followed by underscore and the chain identifier (e.g., 7FAB_L corresponds to chain L of the antibody with PDB code 7FAB).

2.2.3 Similarity calculations

Initially, the anchors of all CDRs of a type (e.g., CDR-L1) were superposed (Kabsch 1976), regardless of length (superposing the anchors reflects how the loops are oriented with respect to the rest of the antibody). To calculate the structural similarity score between CDRs, we used the Dynamic Time Warping (DTW) algorithm (Bellman et al. 1959). The algorithm uses dynamic programming to find the optimum path through

the low-cost areas of a cost matrix (Senin 2008). When two loops of the same length are compared, the algorithm returns the RMSD between the backbone atoms of the loops. When two loops of different lengths are compared, the algorithm calculates the RMSD between backbone atoms of residues matched by the walk through the cost matrix (the method is analogous to the NeedlemanWunsch algorithm for sequence alignment (Needleman et al. 1970), except that the scores are calculated from RMSD between backbone atoms of the residues, instead of being taken from a sequence similarity matrix).

All images of CDR structures were generated using program PyMOL (Schrodinger LLC 2010).

2.2.4 The clustering pipeline

To ensure that the discovered clusters reflect all the underlying structural and sequence patterns, the CDRs were first clustered using the DTW score as a distance measure between structures and the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) (Sokal et al. 1958) algorithm with a cutoff of 1.5 Å. Next, the ability to predict canonical forms from sequence was assessed using Hidden Markov Models (HMM) (see Section 2.2.5). Finally, the canonical forms that contained more than six unique sequences, but could be predicted with less than 75% precision and 25% recall were re-clustered using Density-Based Spatial Clustering of Applications with Noise (DBSCAN) (Ester et al. 1996), choosing the optimal parameter using the Ordering Points to Identify the Clustering Structure (OPTICS) (Ankerst et al. 1999) algorithm (once again using the DTW score as a distance measure). The choice of six sequences was made because the prediction results in smaller clusters were unreliable. The resulting parameters are shown in Table 2.1. This re-clustering with DBSCAN and OPTICS was performed in order to

ensure that every cluster was both structurally coherent and, if the data allowed it, sequence coherent.

In order to ensure there is no drop in accuracy, we have cross-validated our length-independent clustering against a length-dependent version, created using the same methodology, parameters and validation methods.

Using the HMMER predictor, the True Positive Rates (TPRs) and False Positive Rates (FPRs) were calculated across a range of different HMM score thresholds for each cluster.

	Distance cut-off
CDR-L1	0.82 Å
CDR-L2	-
CDR-L3	0.91 Å
CDR-H1	0.80 Å
CDR-H2	0.63 Å

Table 2.1 The parameters for DBSCAN algorithm for each non-H3 CDR type. In the case of CDR-L2 the UPGMA clustering was deemed sufficient.

The TPRs and FPRs were macro-averaged across our clusters and used to plot Receiver Operating Characteristic (ROC) curves, separately for each non-H3 CDR type and, in case of CDR-L1 and CDR-L3, separately for the length-independent and the length-dependent version. To measure the statistical significance of the difference between the length-independent and the length-dependent ROC curves, 1,000 bootstrap replicates were sampled from the TPR and FPR data and the Area Under the Curve was calculated for each ROC replicate. The resulting mean and standard deviation were used to calculate p-values of the difference in AUC. It was found that there is likely no difference between

the curves (the p-values were 0.48 and 0.07 for CDR-L1 curves and CDR-L3 curves respectively).

2.2.5 Cluster prediction from sequence

To predict canonical forms from sequence, the leave-one-out cross-validation procedure was followed. First, the identical CDR sequences were removed from each cluster. Then, one sequence was selected at random and removed from each cluster. Hidden Markov Models (HMMs) were constructed for each cluster from the remaining data using the program HMMER 3.0 (Eddy 1998). Finally, background distribution HMMs were built for each cluster from all sequences outside of the cluster (to use a custom background distribution, rather than the one hardcoded in HMMER, the HMMER source code was modified to return the "raw" log-likelihood rather than the score with the background distribution already subtracted). The selected sequences were scored against the clusters that contained sequences of the same length and assigned to the cluster with which they scored the highest (one-vs-all classification). The procedure was repeated until all sequences had been classified. A similar procedure was followed to score the sequences of loops in clusters containing less than six unique sequences and for loops falling outside the clusters, but in those cases the complete sequence data was used to create HMMs for the large clusters.

To visualize the sequence patterns of the CDR clusters, used as input to our HMMs, we generated sequence logos, using the Weblogo software package (http://weblogo.berkeley.edu/) (Crooks et al. 2004). The sequence logos for the clusters containing at least six unique sequences are shown in the Appendix A1.

2.2.6 Genetic data

Species and germline data were extracted from the IMGT database (International ImMunoGeneTics information system®, http://www.imgt.org) (Lefranc et al. 1999) and from the SAbDab (Dunbar et al. 2014) database when the respective IMGT entry was not available. If there was a discrepancy between the species annotation in the IMGT record and the PDB file header, or if a human germline was reported for a CDR belonging to a cluster containing primarily mouse antibodies (or vice versa), the article associated with the PDB entry was inspected to learn the origin of the CDRs.

2.3 Results & discussion

The structures of CDR loops were extracted from antibody structures available in the SAbDab (Dunbar et al. 2014) database and filtered as described in the methods Section 2.2.2).

Using the structural alignment produced by the dynamic time warping (DTW) algorithm we found that across all CDR types, in about 50% of cases the insertion site identified by Chothia alignment is structurally correct and in about 77% of cases the correct site is within one residue of the Chothia site.

Taking all the unique CDR sequences from our structural set, we identified the structurally closest loop to each using the dynamic time warping (DTW) score (see methods Section 2.2.3). In all CDR types, apart from CDR-L2, for some fraction of CDRs the structurally closest partner was of a different length (Table 2.2, Figure 2.1). This result suggested that length-independent canonical classes could exist.

CDR type	CDRL1	CDRL2	CDRL3	CDRH1	CDRH2	CDRH3
Number of structures	1701	1762	1752	1734	1779	1671
Number of unique sequences	455	302	518	374	493	614
Number of times the closest structure is of a different length	20	0	35	18	15	288
Fraction	4%	0%	7%	5%	3%	47%

Table 2.2 Length-independent structural similarity. For each CDR type the table shows: First row - number of CDR structures, after the filtering described in the methods Section 2.2.2 was applied. This is also the number of structures that were used as input to our clustering method. Second row - number of unique CDR sequences. Third row - number of unique sequences for which the closest structural partner is of a different length. Fourth row - fraction of unique sequences for which the closest structural partner is of a different length.

Motivated by this result, we combined ideas from density-based and hierarchical clustering methods to create length-independent canonical classes. We used all CDR structures, regardless of sequence redundancy, as input to our clustering method (see methods Section 2.2.4). Using the length-independent methodology, we discovered 26 large clusters in total, four of which contained CDRs of more than one length (for a cluster to be classified as large it had to contain at least six unique sequences). The results for the large clusters are summarized in Table 2.3.

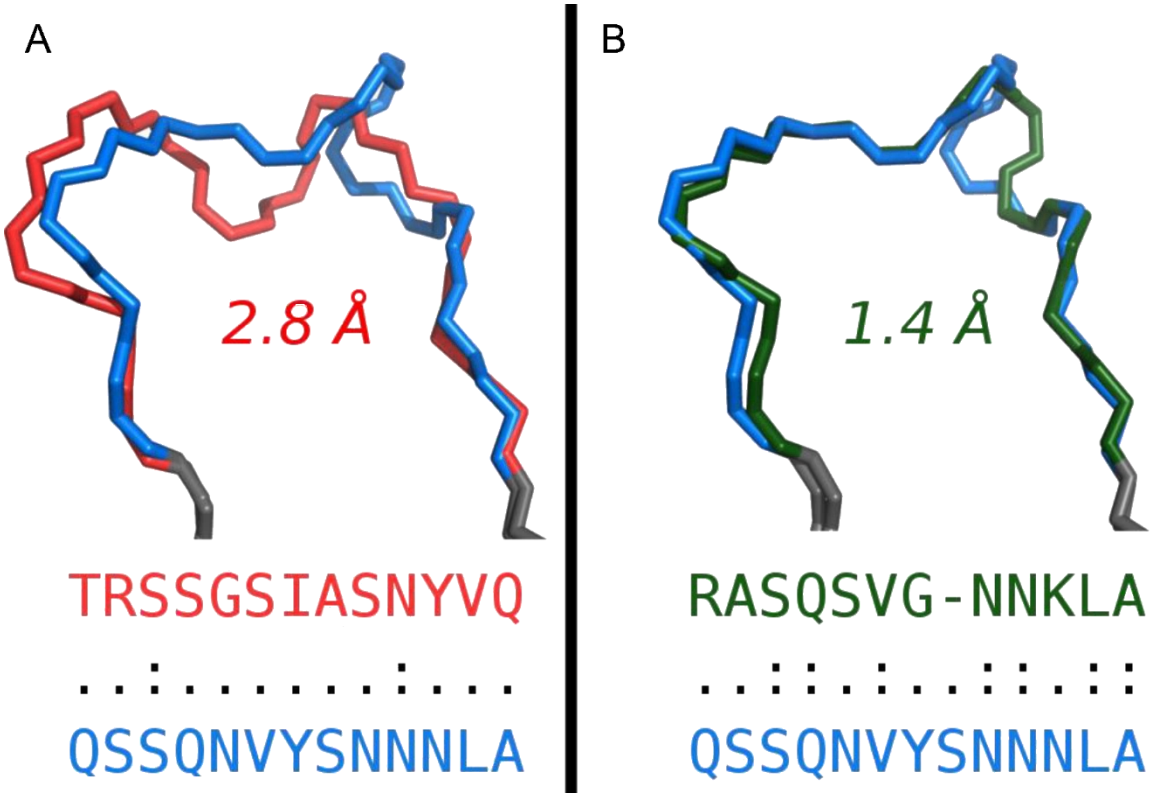

Figure 2.1 A) Structure of CDR-L1 from 4JO2_M (blue, length 13) aligned with its closest structural partner of the same length, the CDR-L1 from 3BDX_A (red, length 13), which is 2.8 Å away, as measured using the DTW score. The loops have only two identical residues. B) Structure of CDR-L1 from 4JO2_M (blue, length 13) aligned with its closest structural partner of different length, the CDR-L1 from 3LHP_M (green, length 12), which is 1.4 Å away, as measured using the DTW score. The loops have seven residues in common. In both panels A and B the anchors of the CDRs are shown in grey.

We find that most of the large light chain clusters contain only either the κ or λ light chains. The two exceptions are L3-5-A and L3-9-A. The cluster L3-9-A has been described previously by North et al. (as the cluster L3-9-1). The cluster L3-5-A contains structures that were not available at the time the work of North et al. was published and

are all from broadly neutralizing antibodies, suggesting that such loops tend to take a similar shape, irrespective of the chain type.

We use the following nomenclature for our clusters: two letters describing the CDR type, followed by a dash and the lengths of the CDRs contained within the cluster, separated by commas, followed by another dash and a capital letter describing the order of the cluster (e.g., L1-13,14-A corresponds to the first cluster containing CDR-L1 structures of lengths 13 and 14).

2.3.1 Clustering details

Here we describe in detail the clustering results for each non-H3 CDR type. We include descriptions of interactions that stabilize the loop structures and describe sequence patterns. The clusters are ordered first by length of the shortest loop, then by number of structures and, finally, by number of sequences.

2.3.1.1 CDR-L1 clusters

CDR-L1 is the second most length variable of all CDR types, behind CDR-H3. The shortest loops of this type are of length seven and the longest of length 17. There are 17 clusters in total, 10 containing at least six sequences, two of which contain loops of more than one length.

The cluster L1-10,11,12-A contains the majority of loops of length 10 and 11 and three loops of length 12 (see Figure 2.2A).

Cluster name	Length	Number of structures	Middle structure	Number of unique sequences
CDR-L1 (κ)				
L1-10,11,12-A	10, 11, 12	779	3SOB_L	204
L1-12-A	12	22	1HQ4_A	12
L1-15-A	15	55	3QRG_L	26
L1-16-A	16	273	1KFA_M	65
L1-17-A	17	113	2R1X_A	31
CDR-L1 (λ)				
L1-11-A	11	38	4IMK_C	9
L1-11-B	11	24	3MLS_M	8
L1-13,14-A	13, 14	117	4FQJ_L	37
L1-13-A	13	23	2WOL_C	6
L1-14-A	14	92	1YOL_C	7
CDR-L2				
L2-7-A	7	1708	2G5B_A	291
L2-7-B	7	21	3I9G_L	6
CDR-L3 (mixed λ and κ)				
L3-5-A	5	17	4JPI_B	6
L3-9-A	9	107	1Y0L_C	22
CDR-L3 (κ)				
L3-8-A	8	106	4HGW A	29
L3-9,10-A	9, 10	1133	3RVV_C	335
CDR-L3 (λ)				
L3-10,11-A	10, 11	53	3MLX_L	23
CDR-H1				
H1-7-A	7	1267	1PLG_H	357
H1-7-B	7	18	4FQQ_F	6
H1-8-A	8	37	3RVW_D	8
H1-9-A	9	86	3IDN_B	9
CDR-H2				
H2-7-A	7	387	3ZKM_H	91
H2-8-A	8	650	1I8M_B	197
H2-8-B	8	305	2VXS_K	93
H2-8-D	8	19	1YQV_H	9
H2-10-A	10	147	3HZV_B	25

Table 2.3 Information on CDR clusters that contain at least six unique sequences. The following nomenclature is used: two letters describing the CDR type, followed by a dash and the lengths of the CDRs contained within the cluster, separated by commas, followed by another dash and a capital letter describing the order of the cluster (e.g., L1-13,14-A corresponds to the first cluster containing CDR-L1 structures of lengths 13 and 14). The "middle structure" column shows the PDB id and the name of the chain containing the CDR structure that is in the centre of the corresponding cluster. The clusters are ordered first by length, then by number of structures and finally by number of sequences.

The most common interaction, present in 194 out of 204 sequences in this cluster, is the hydrogen bond formed between a Ser at Chothia position 26 and the backbone of the residue at position 3. In addition, all CDRs in this cluster contain a hydrophobic residue at Chothia position 33, the side chain of which is buried inside the loop structure.

The second largest cluster of CDR-L1s of length 11 is L1-11-A, coded for by germlines from the lambda Immunoglobulin locus. Loops in this cluster do not have a Ser at Chothia position 26; instead all loops in this cluster display an interaction between residues at Chothia positions 27 and 30 creating a more "compact" structure, aided by presence of two Gly (present in all loops in this cluster) at Chothia positions 25 and 29.

The third cluster containing loops of length 11 - L1-11-B contains structures which in all cases are stabilized by a hydrogen bond between the backbone oxygen of Leu at Chothia position 28 and the backbone nitrogens of Chothia residues 25 and 31.

Many of the length 12 CDR-L1s are found in cluster L1-12-A which consists of 12 unique sequences. In 11 of the 12 unique sequences in this cluster, there is a hydrogen bond between the sidechain of the residue at Chothia position 30 and the backbone of Chothia residue 31, an interaction not present in L1-10,11,12-A.

Most unique sequences of length 13 and 14 belong to cluster L1-13,14-A, which forms a compact shape stabilized by a number of interactions. All of the CDRs in this cluster form hydrogen bonds between the backbones of Chothia residues 26 and 29, 26 and 30A, and 29 and 30B (Figure 2.2B). The loops in clusters L1-13-A and L1-14-A, which contain the remaining structures of length 13 and 14, respectively, lack the aforementioned interactions which results in a more extended shape without the "hoop" in the centre of the CDR.

The majority of the longer CDR-L1 loops are contained in clusters L1-15-A, L1-16-A and L1-17-A. The first part of long CDR-L1 loops, up to the central hydrophobic residue at Chothia position 29, resembles the loops in cluster L1-10,11,12-A. The remaining part of the loop structure is a protrusion of varying length stabilized by hydrogen bonds between backbone atoms (see Figure 2.2C).

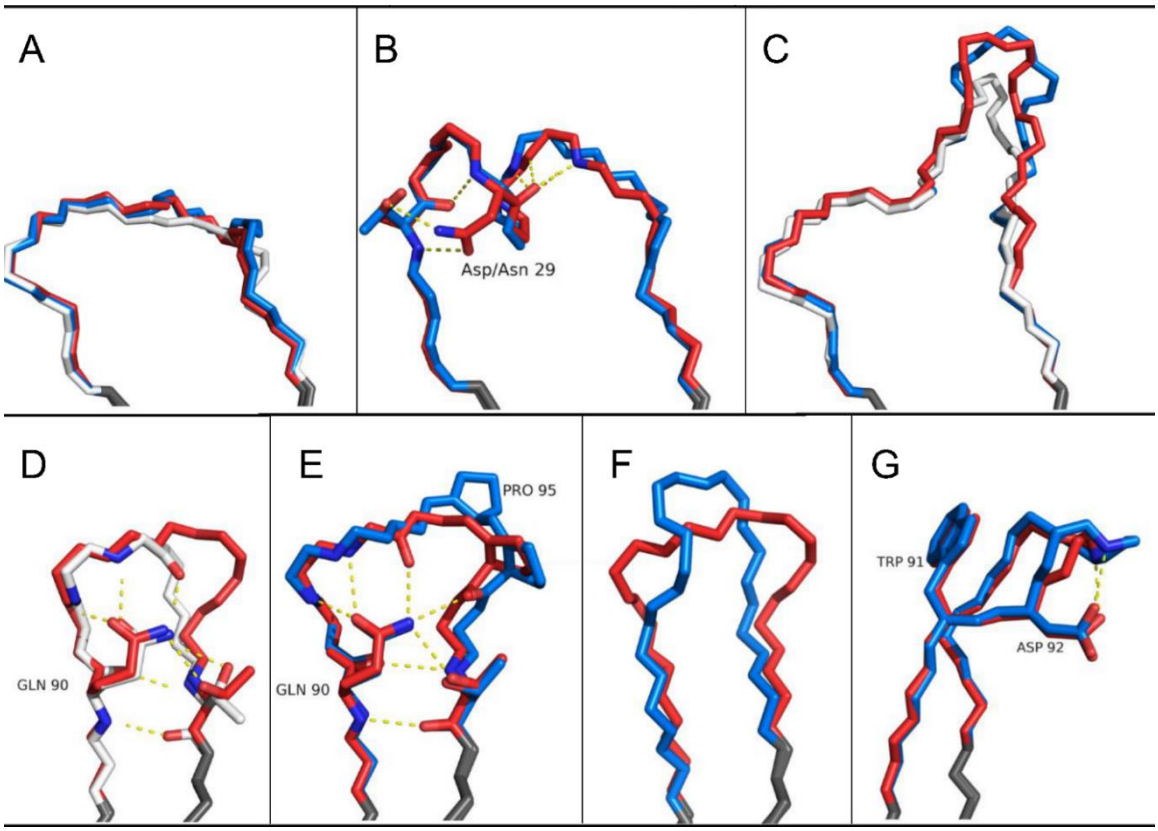

Figure 2.2 CDRs with different lengths, but similar structures, with their anchors aligned, shown in grey. A) CDR-L1 of 4F33_E (length 10, white), 3SOB_L (length 11, red) and 3EO0_A (length 12, blue) from cluster L1-10,11,12-A. The three CDRs are structurally similar, except for the insertion site at position 30. B) CDR-L1 loops of 4FQJ_L (length 13, red) and 3U2S_L (length 14, blue) from cluster L1-13,14-A. Presence of Asp/Asn at position 29 (the side chain of which is shown in the figure, along with the hydrogen bonds it makes) creates a "hoop" in the structures of both lengths in cluster L1-13,14-A. C) CDR-L1 of 3QRG L (length 15, white, cluster L1-15-A), 1KFA M (length 16, red, cluster L1-16-A) and 2R1X_ A (length 17, blue, cluster L1-17-A). The structures of longer CDR-L1 loops containing a "protrusion" starting around residue 29. D) CDR-L3 of 4HGW_A, (length 8, white, cluster L3-8-A) and 3RVV_C, (length 9, red, cluster L3-9,10-A). The conformations of CDRs in both clusters appear to be maintained by Gln at position 90 (the side chain of which is shown, along with the contacts to other residues). E) CDR-L3 of 3RVV_C, (length 9, red) and 4HHA_A, (length 10, blue) from cluster L3-9,10-A. CDRs of both length have similar conformations, maintained by the Gln at position 90, with the CDR of length

10 containing an additional Pro at position 95A (The side chains of both Pro are shown in the figure). F) CDR-L3 of 3RVV_C (length 9, red, cluster L3-9,10-A) and 1Y0L_C (length 9, blue, cluster L3-9-A). The alignment shows the difference in conformation between the two largest clusters containing CDRs of length 9. G) CDR-L3 of 3MLX L, (length 10, red) and 4NZT_L, (length 11, blue) from cluster L3-10,11-A. The conformation of both CDRs seems to be stabilized by the interaction between Asn at position 92 and the backbone nitrogen of residue 94 and the presence of large hydrophobic residue at position 91 (the side chains of Asn-92 and Trp-91 are shown in the figure).

2.3.1.2 CDR-L2 clusters

CDR-L2 is the least variable in terms of length of all CDRs, having only two lengths in our set - seven or 11. There are five clusters out of which only two contain more than six unique sequences (see Appendix A2).

The vast majority of the CDR-L2s of length seven are in cluster L2-7-A. This shape occurs universally across all available species and in both κ and λ light chains. The second largest cluster is L2-7-B, also containing loops of length seven. The structures of the CDRs in the two clusters differ only by the conformation of the first three residues. The two clusters containing CDR-L2s of length 11, L2-11-A and L2-11-B, are very small and contain only two and three sequences respectively. L2-11-B is more diverse of the two containing two loops from Rhesus Monkey anti-HIV antibodies and a CDR-L2 from a human pre B-cell receptor.

2.3.1.3 CDR-L3 clusters

CDR-L3 loops in our set span lengths between five and 12 residues. The majority of CDR-L3 loops are of length nine. There are 11 clusters in total, five containing at least six unique sequences, two of which contain loops of more than one length (see Appendix A2).

All CDR-L3 loops of length five are found in cluster L3-5-A and are all from HIV-1 neutralizing antibodies that recognize the epitope of a CD4 binding site (Wu et al. 2011).

The cluster L3-8-A contains the majority of loops of length eight. The structure of these loops is stabilized by Gln at Chothia position 90 forming hydrogen bonds with residues at positions 92 (in 97 out of 106 cases), 93 (in 91 out of 106 cases) and the side-chain of the Thr at position 97 (in 86 out of 106 cases) (Figure 2.2D).

L3-9,10-A contains the largest number of CDR-L3 loops and contains loops of length nine and ten. Most loops of length nine belong in this cluster. The loops in this cluster are stabilized by the same interactions as the loops in cluster L3-8-A the difference being that 325 out of 335 of loops of length nine contain a Pro in cis conformation at Chothia position 95 and all loops of length ten contain two Pro at Chothia positions 95 and 95A. The Pro create a sharp turn making the loop conformations similar for loops of both lengths (Figure 2.2E).

Cluster L3-9-A is the second largest cluster containing loops of length nine (107 structures, 22 sequences). The residue at Chothia position 91 usually contains a large side-chain with an aromatic ring (Trp, Tyr, Phe). The CDR conformation is stabilized, in 20 out of 22 cases, by a hydrogen bond between backbone carboxyl oxygen of Chothia residue 92 and backbone nitrogen of Chothia residue 94.

Figure 2.2F illustrates the differences in conformation between CDR-L3 loops of length nine in cluster L3-9,10-A and in cluster L3-9-A.

The structures of CDR-L3 loops that are part of the cluster L3-10,11-A are stabilized by similar interactions to the CDRs in cluster L3-9-A. The difference is the presence of a highly-conserved Asp/Asn forming a hydrogen bond with the residue at Chothia

position 94 creating a bend in the structure (Figure 2.2G). In all CDRs in this cluster there is a hydrogen bond between backbone nitrogen of residue at Chothia position 92 and the backbone oxygen of residue at Chothia position 95. In 17 out of 23 sequences, there is also a hydrogen bond between backbone nitrogen at Chothia position 93 and the backbone carboxyl oxygen of Chothia residue 28, which is a part of the CDR-L1 loop.

2.3.1.4 CDR-H1 clusters

CDR-H1 loops are three to 13 residues long. CDRs of length seven are most prevalent (87% of all structures). There are 14 clusters, but out of these only four contain at least six unique sequences. There are no clusters containing loops of more than one length. Overall, this CDR type contains the largest number of small clusters, mostly coming from camelid antibodies (see Appendix A2).

The largest cluster of CDR-H1 is H1-7-A containing 73% of all H1 structures. CDRs in this cluster tend to include (221 out of 257 sequences) a residue with a large aromatic ring at Chothia position 27 (Phe or Tyr) and, in 218 out of 257 of cases, a hydrogen bond between backbone atoms of Chothia residues 28 and 31.

The second cluster containing CDRs of length seven is H1-7-B. The difference in conformation between CDR-H1 loops in H1-7-A and H1-7-B is subtle, but most pronounced around Chothia residues 28 and 31.

The other two large clusters are H1-8-A and H1-9-A, containing CDRs of length eight and 10, respectively. All structures in cluster H1-8-A contain a hydrogen bond between the backbone atoms of Chothia residues 28 and 31. The structures of loops in cluster H1-9-A are similar to H1-8-A, except for a large "bulge" around Chothia residue 32. All CDR structures in cluster H1-9-A contain a hydrogen bond between backbone atoms of Chothia residues 31 and 32.

2.3.1.5 CDR-H2 clusters

CDR-H2 loops in our set show only six lengths, between seven and 12 residues. Most of the CDRs of this type are of length seven, eight or 10. There are 12 clusters in total, five containing at least six unique sequences (see Appendix A2). There are no clusters containing loops of more than one length but the conformations in clusters H2-7-A (containing CDRs of length seven) and H2-8-A (containing CDRs of length eight) are similar.

The majority of CDR-H2s of length seven belong to cluster H2-7-A. The sequence pattern for this cluster shows a preference for Gly at Chothia position 55 and a residue with side-chain oxygen at Chothia position 56 (Ser/Thr/Asn/Asp).

There are two clusters containing loops of length eight - H2-8-A and H2-8-B. In cluster H2-8-A, 195 out of 197 structures contain a hydrogen bond between backbone atoms of Chothia residues 52 and 55 (same as H2-7-A). In addition, 154 loops contain a Pro at Chothia position 52A and a Gly at Chothia position 56. The structures of CDRs in this cluster are similar to H2-7-A with the Pro-52A creating a sharp turn in the loop.

In contrast, the CDR structures in cluster H2-8-B lack the Pro-52A and the hydrogen bond between backbone atoms of Chothia residues 52 and 56. Instead, 92 out of 93 structures contain a hydrogen bond between Chothia residues 52 and 54, and, in 76 cases a hydrogen bond between side-chain of Chothia residue 52 and Chothia residue 56. In 84 cases, the conformation in cluster H2-8-B is also stabilized by the hydrogen bond between the side-chain nitrogen of the framework residue Arg-71 and the backbone oxygen of residue at position 52.

Most of the CDR-H2 loops of length 10 are concentrated in cluster H2-10-A. Loops of this length are usually from mouse antibodies. All structures in this clusters contain

hydrogen bonds between residues 52A and 55 and 52B and 54. The structures of loops in this cluster in 23 out of 25 cases also display the interaction between Arg-71 and the backbone carboxyl oxygen of residue at Chothia position 52.

It is possible for CDRs of the same sequence to belong to different clusters. We observed this structural variation is most common in CDR-H2. For example, the CDR-H2 loop with sequence EILPGSGS in the unliganded structure 1MLB_B belongs to cluster H2-8-A but in the bound form, in structures 1MLC_B and 1MLC_D, the loop belongs to cluster H2-8-D with 2.1 Å difference in RMSD (1MLB are 1MLC are crystal structures of the same antibody). YIWPSGGN is the sequence of CDR-H2 loops in structures 3HI6_X and 3HI6_H and in 3HI6_X the loop belongs to cluster H2-8-B but in 3HI6_H the loop is part of cluster H2-8-H. This result implies there are CDRs which can exist in different structural states. Unfortunately, it will also confuse cluster prediction from sequence.

Previous analyses of CDR structures (Tramontano et al. 1990; North et al. 2011) discussed how the framework residue at Chothia position 71 influences the conformation of CDR-H2. We analysed the amino acid distribution of residue 71 across our largest clusters and found that the framework sequences in clusters H2-8-B and H2-10-A show a clear preference for Arg at this position, in agreement with the previous work.(Tramontano et al. 1990; North et al. 2011) We also find that, compared to previous work, the framework sequences in cluster H2-8-A show an increased abundance of Arg at position 71 (~5% in equivalent North et al. cluster H2-10-1, ~10% in H2-8-A), making the residue less predictive of cluster membership (see Figure 2.3).

Having described the clustering results in detail, in the next section we focus specifically on the clusters containing loops of more than one length.

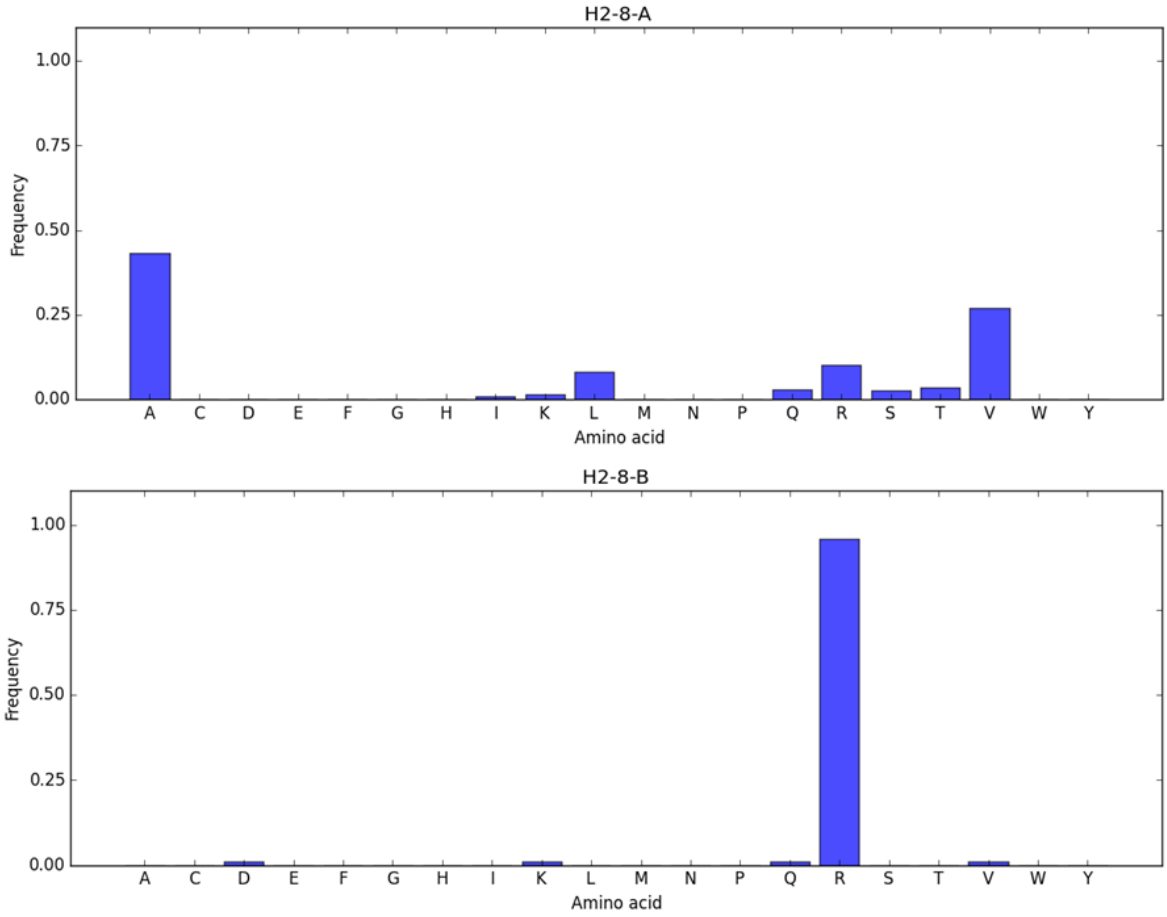

Figure 2.3 The amino acid distributions of the framework residue at Chothia position 71, plotted for two clusters of CDR-H2 loops of length 8. The y-axis in each panel shows the frequency of each amino acid type and the x-axis shows all standard amino acids.

2.3.2 Sequence patterns in length-independent clusters

For the concept of length-independent structural similarity to be useful in loop modelling, the structural relationships between CDRs of different length must be matched by sequence similarity. To investigate whether the length-independent clusters contain clear sequence patterns, we compared the performance of the HMM prediction method (see methods Section 2.2.4) to the length-dependent version of our clustering. We find that the increased number of sequences in the length-independent clusters improves the precision of prediction. Figure 2.4 illustrates this principle with the example

of CDR-L1 cluster L1-13,14-A, which contains λ CDRs of length 13 and 14. If the cluster is split by length, prediction precision decreases. There are clear similarities between the sequence logos of CDRs of length 13 and length 14, especially the presence of Asn/Asp at Chothia position 29, which appears to be key for maintaining the structures of the loops in this cluster.

The importance of consistent sequence patterns is further illustrated by the CDR-L3s of length 10, which are part of the cluster L3-10,11-A. These CDRs have no close structural homologs among the other CDR-L3s of length 10 and, in the length-dependent version of the clustering, are not clustered. In the length-independent version of the clustering, they are part of the cluster L3-10,11-A, which contains primarily CDRs of length 11.

To assess the global performance of the prediction method on our clusters, we plotted receiver operating characteristic curves for each CDR type (see Figure 2.5). The area under the curve (AUC) for each CDR type was above 0.90 (a perfect model would get an AUC score of 1 while a random predictor would receive a score of 0.5).

We show in the next section how our clustering improves predictions in the context of next-generation sequencing (NGS) of CDR-L3 repertoire.

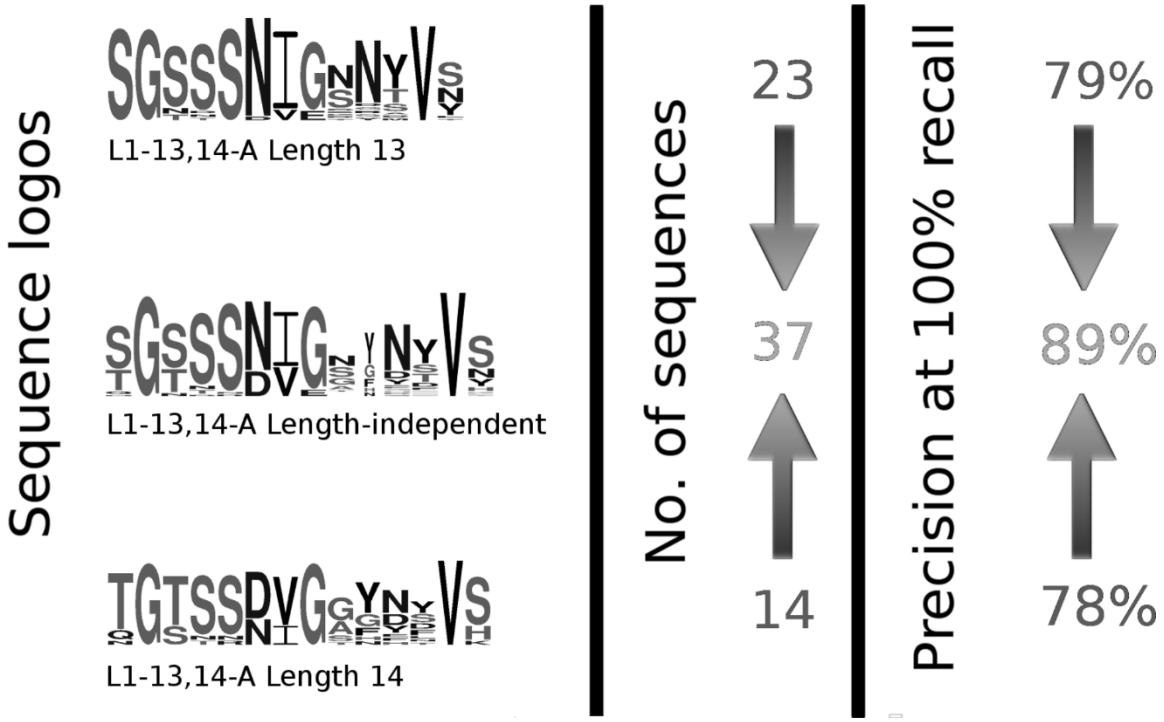

Figure 2.4 An illustration of how length-independent clustering improves the precision of prediction. The first column shows logos created using sequences of CDRs of length 13 (top) and 14 (bottom) inside cluster L1-13,14-A, with the logo for the complete length-independent cluster in the middle. The second column shows the number of sequences of each length (top and bottom) and the number of sequences in the complete length-independent cluster (middle). In the third column, the precision at 100% recall is reported for the complete cluster (middle) and for the two length-dependent clusters resulting from splitting L1-13,14-A by length (top and bottom). All cluster logos in this thesis were created using the Weblogo software (Crooks et al. 2004) (http://weblogo.berkeley.edu/).

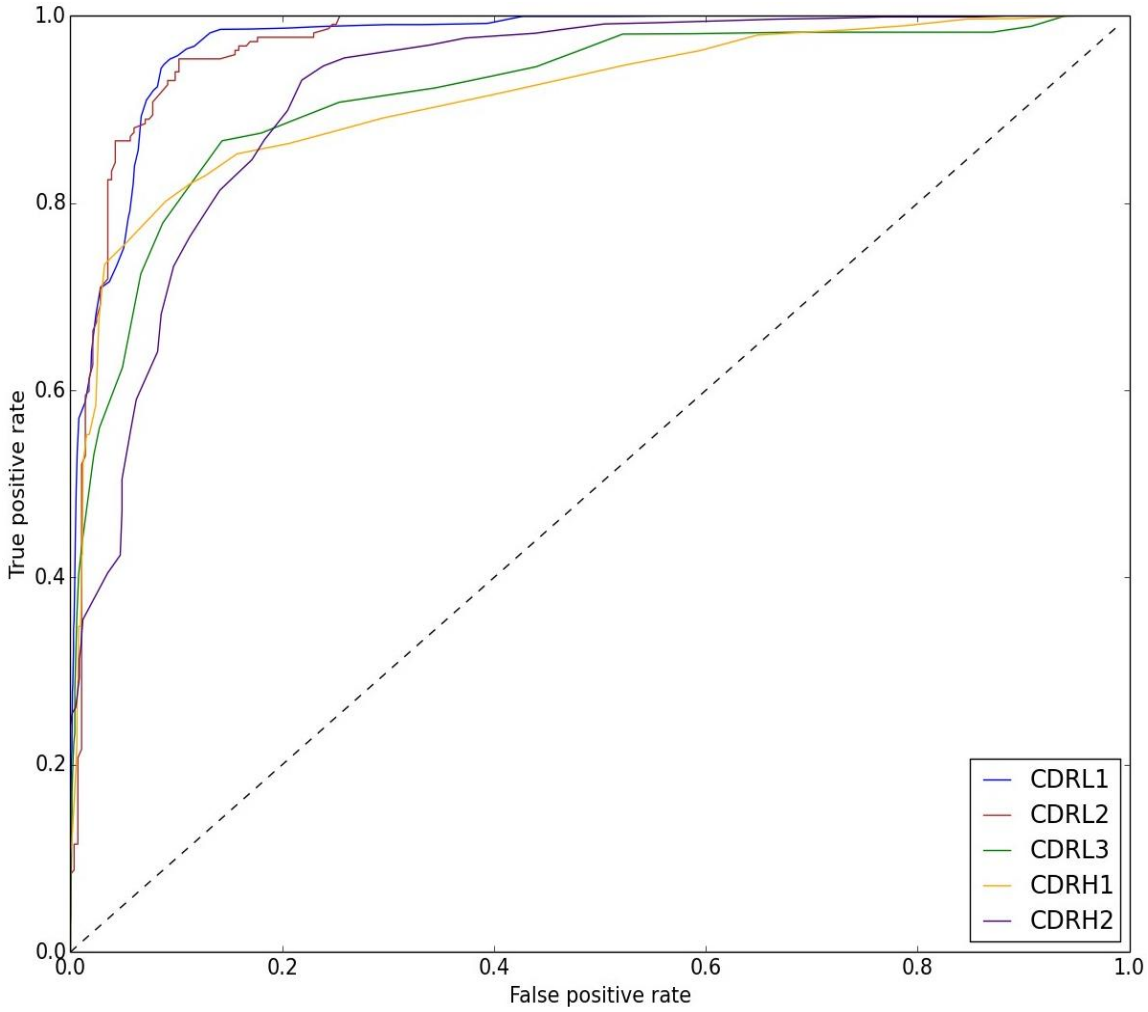

Figure 2.5 Receiver Operating Characteristic (ROC) curves for length-independent clustering for all CDR types, obtained by macro-averaging the results for each constituent cluster for each CDR type. The ROC curve for CDRL1 is shown in blue, for CDRL2 in brown, for CDRL3 in green, for CDRH1 in yellow and for CDRH2 in violet. The Area Under the Curve (AUC) for CDRL1 is 0.97, for CDRL2 is 0.97, for CDRL3 is 0.92, for CDRH1 is 0.91 and for CDRH2 is 0.92. A perfect model would get an AUC score of 1.0 while a random predictor would receive a score of 0.5.

2.3.3 Analysis of next-generation sequencing data.

As the length-independent clusters contain such clear sequence patterns, making them useful for prediction, we investigated whether the small gains in prediction coverage shown in the structural set have a significant effect when considering the large Next-

Generation Sequencing (NGS) sets of CDR-L3 sequences. We examined three large antibody NGS datasets: the first dataset was created through sequencing experiments performed by UCB Pharma Ltd and contains over ~9,000,000 human light chain sequences; the second dataset was obtained by (DeKosky et al. 2014) and contains 198,148 human paired CDR-H3 - CDR-L3 sequences from three donors; and the third dataset was extracted from the DIGIT database (Chailyan et al. 2012) and consists of 71,404 light chain sequences from over 100 different species. Since only the CDR-L3 sequences were available in all datasets, we extracted the unique sequences of this type, obtaining ~1,000,000 sequences from the UCB dataset, 72,045 from the DeKosky et al. dataset and 12,960 from the DIGIT dataset.

We found that the length-distribution of CDR-L3 sequences in these datasets differs significantly from the length distribution of CDR-L3s whose structure is known (see Figure 2.6). For example, sequences of length 10 comprise ~26% of the UCB dataset (290,000 sequences) and only ~6% of the SAbDab database. A major reason for this disparity is the relative abundance of κ chains in the structural dataset in comparison to the NGS dataset. The structural dataset consists of about 78% κ Light chains and 22% λ Light chains while a more balanced distribution of 47% κ chains and 53% λ chains is observed in the NGS dataset (which contains only human sequences). Nevertheless, even after separating the CDR-L3 sequences by the chain type we still observe that the sequences of length nine are overrepresented and sequences of length 10 underrepresented in the structural dataset (see Figure 2.7). Due to this disparity, the canonical class assignment would be more difficult if performed in a length-dependent way.

To test whether we can assign more sequences to clusters using the length-independent methodology, we evaluated the cluster membership of the unique CDR-L3 sequences in both a length-dependent and length-independent way at expected precisions between 75% and 90% (see Figure 2.8). Precision of cluster membership assignment was estimated using the structural data and the HMM scores returned by HMMER (Eddy 1998) (see methods Section 2.2.5). We found that across all three datasets we can predict more sequences using the length-independent approach. For example, at 80% precision, we can assign into clusters an additional ~125,000 sequences (~21% improvement, Figure 2.8A) from the UCB dataset, 8,958 sequences (~21% improvement, Figure 2.8B) from the DeKosky et al. dataset and 1,338 sequences (~17% improvement, Figure 2.8C) from the DIGIT dataset.

Together, these results illustrate that using length-independent clustering we can structurally characterize a much larger part of antibody sequence space.

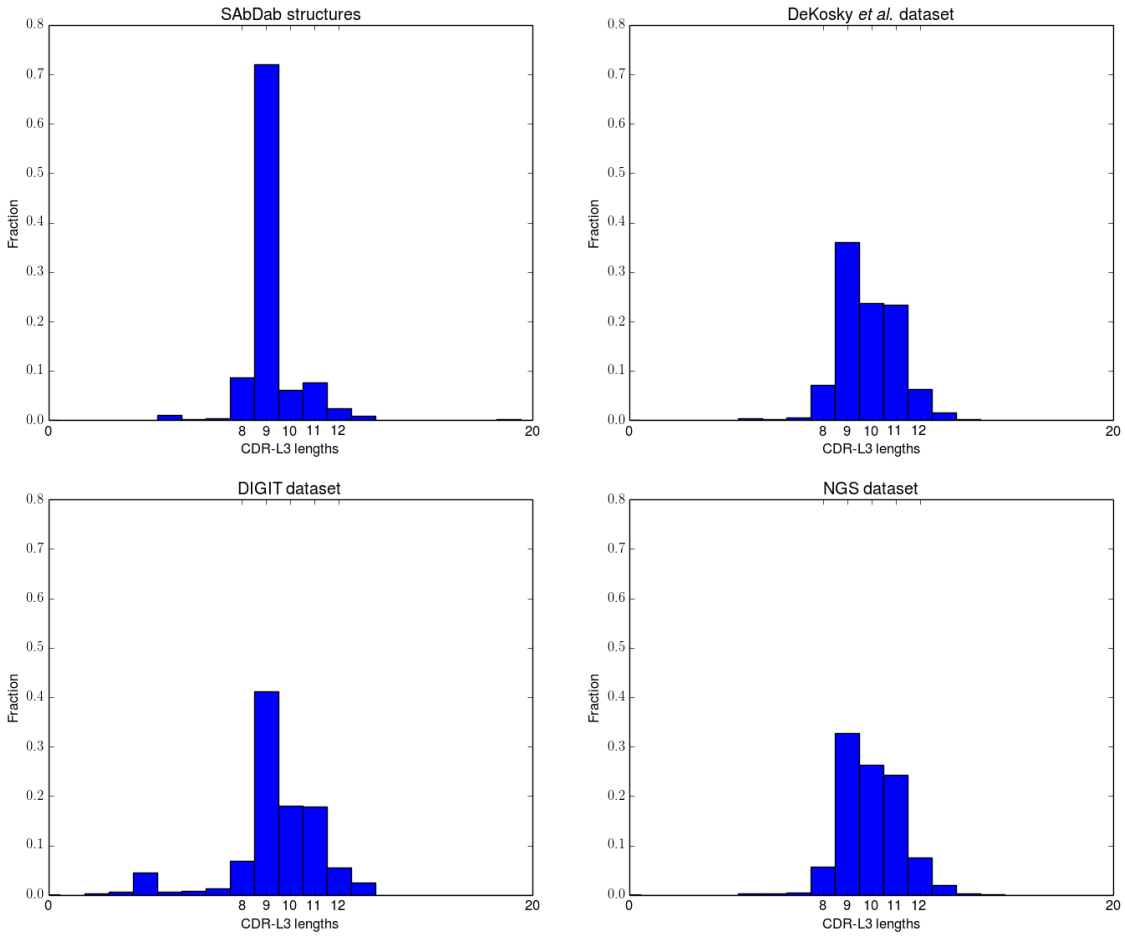

Figure 2.6 Comparison between the length distributions of the unique CDR-L3 sequences in the three NGS datasets and the distribution in our structure data. The top left panel shows the distribution in our structural dataset, the top right panel shows the distribution in the (DeKosky et al. 2014) dataset, the bottom left panel shows the distribution for the DIGIT (Chailyan et al. 2012) dataset and the bottom right panel shows the distribution in the UCB NGS dataset. The histograms have been normalized and the height of the bars show the fraction of sequences with a particular length. The distributions of lengths in the UCB NGS and the DeKosky et al. datasets are similar, as they both contain only sequences of human origin. The structural dataset contains mostly sequences of length nine, other lengths are significantly underrepresented.

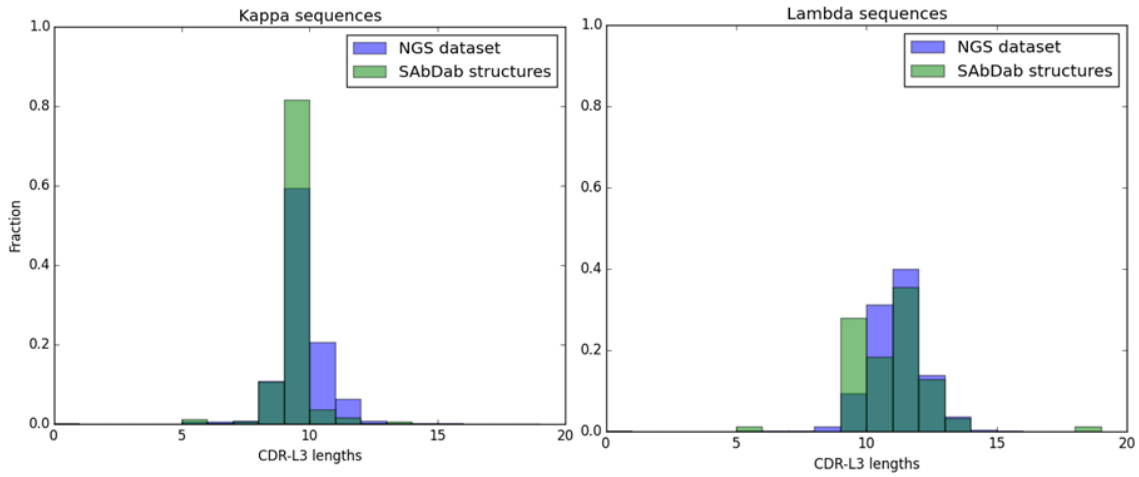

Figure 2.7 Comparison between the distribution of lengths of unique CDR-L3 sequences in the UCB NGS dataset and in our structural data. The NGS distribution is shown in blue and the structural distribution is shown in green. The histograms have been normalized and the height of the bars shows the fraction of sequences that have a particular length. The graph shows that CDR-L3 sequences of length nine are overrepresented and the sequences of length 10 are underrepresented in the structural dataset for both kappa and lambda chains.

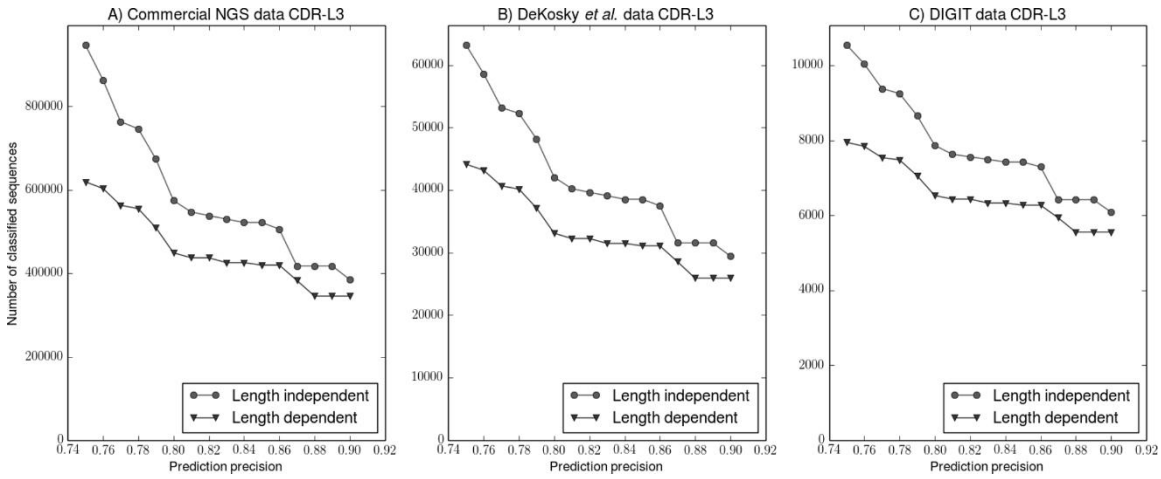

Figure 2.8 Length-independent clusters increase the number of sequences that can be classified. The expected precision of prediction (x axis) was calculated from our structural data based on the HMM score returned by HMMER. The circles show the number of sequences that can be classified using our length-independent approach, while the triangles show the number of sequences that can be classified by the length-dependent approach. A) The classification of ~1,000,000 unique CDR-L3 sequences from the UCB dataset. At 0.8 precision, we can classify 125,000 or about 21% more sequences into clusters. B) The classification of 72,045 CDR-L3 sequences from the (DeKosky et al. 2014) dataset. At 0.8 precision, we can classify 8,958 or ~21% more sequences into clusters. C) The classification of 12,960 CDR-L3 sequences from the DIGIT (Chailyan et al. 2012) dataset. At 0.8 precision, we can classify 1,338 or ~17% more CDR-L3 sequences into clusters.

2.3.4 Reasons for length-independent structure similarity

Because our length-independent clusters show strong sequence patterns, we investigated the possible causes of similarity between CDR structures of different lengths. We propose three natural mechanisms for the generation of structurally and sequence similar CDRs of different lengths.

Firstly, the germline contains a large repertoire of V-region genes (Retter et al. 2005). One of the causes of similarity between structures of different lengths appears to be

the identity of certain key residues, common between different germlines (see Figure 2.9A).

Secondly, in the early stages of development the antibody-producing B cells undergo a somatic recombination, during which V (variable), J (joining) and, in the case of the heavy chain, D (diversity) gene segments are spliced together (see Section 1.3.4.1). This results in a novel sequence for the variable domain of the antibody. The VJ recombination affects the sequence of CDR-L3, which explains why CDR-L3 is more variable than the other light chain CDR types (Tonegawa 1983). We have found that the different rearrangements of the V and J genes may not always result in a significant change to the CDR structure, which could lead to shape similarity between CDR-L3 loops of different lengths (Figure 2.9B).

Thirdly, B cells proliferate when they are stimulated by antigens. During this proliferation, the V-region coding sequences of both heavy and light chain accumulate point mutations at a rate that is about a million times greater than in other genes (Teng et al. 2007). The few mutated B cells, which express antibodies with higher affinity, are further stimulated to proliferate. This process, known as somatic hypermutation, can result in a 1000-fold increase in affinity to the target (Li et al. 2004). During the hypermutation phase, deletions and insertions may arise, although they are far less common than substitutions (Wilson et al. 1998; Briney et al. 2012). The change in sequence length generated by somatic hypermutation may result in two CDRs having similar structure, despite being of different length. A possible example of this is shown in Figure 2.9C.

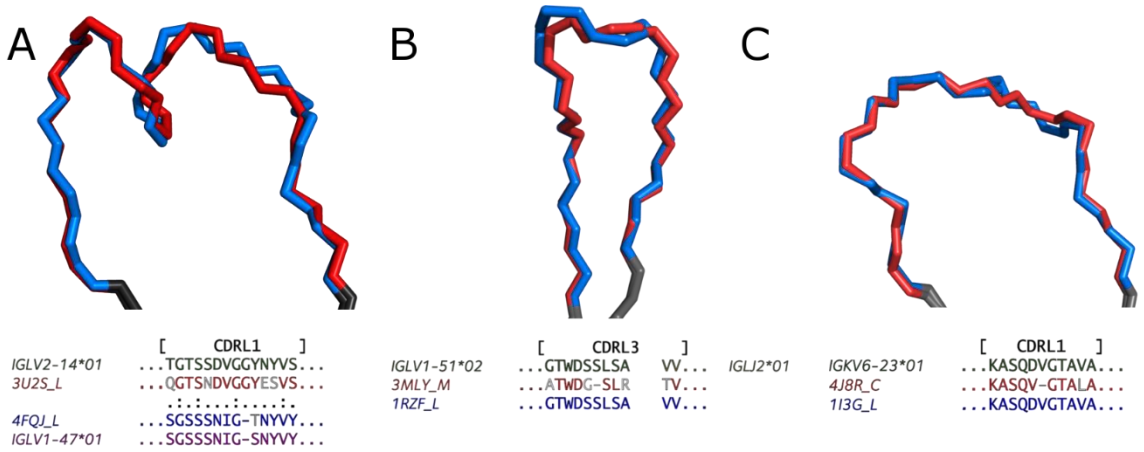

Figure 2.9 CDRs with different lengths, but similar structures, with their anchors aligned, shown in grey. This figure demonstrates how length-independent shape similarity may arise. A) CDR-L1 of 3U2S_L (length 13, red) and 4FQJ_L (length 14, blue). The two CDRs are coded for by human germlines from different subgroups (IGLV2-14*01 and IGLV1-47*01 respectively), but the identity of certain key residues results in a similar shape. Especially important seems to be the presence of Asp/Asn at Chothia position 29. B) CDR-L3 of 3MLY_M (length 10, red) and 1RZF_L (length 11, blue). The two CDRs have similar structures and appear to be coded for by the same human V-gene (IGLV1-51*02) and human J-gene (IGLJ2*01). The observed length difference seems to be caused by different rearrangement of genes during VJ recombination. C) CDR-L1 of 4J8R_C (length 10, red) and 1I3G_L (length 11, blue). This is an example of two structurally similar CDRs that appear to come from the same murine germline (IGKV6-23*01), but in the case of 4J8R_C an Asp has been deleted during somatic hypermutation.

Assuming that the human germline repertoire contains ~40 functional variable genes of each type (heavy, λ, κ), five functional joining genes of each type, 23 functional diversity genes, and that the N-diversity and somatic hypermutation increase the number of possible Light and Heavy chain sequences by about 1000-fold, we can estimate that the human organism can produce about 10^{11} distinct antibodies. The fact that we observe length-independent structural similarities in the limited number of antibody crystal

structures available to us suggests that it may be a relatively common occurrence in nature.

2.3.5 Comparison to previous clusterings

As we noted above, many length-dependent clusterings of CDR structures have previously been reported. In this section, we describe the differences between our clustering and a recent clustering of CDRs into length-dependent canonical classes by North et al. (2011). Tables containing the full comparison are given in Appendix A3.

The large clusters (those containing at least six unique sequences) map well from our work to North et al. (2011), usually having one-to-one correspondence. Some clusters, however, are split or joined due to differences in methodology or length-independence. For example, loops of length 11 from our cluster L1-10,11,12-A are split into two clusters L1-11-1 and L1-11-2 in the work of North et al. This cluster is split by North et al. due to a change in conformation of a single residue at position 30. This does not lead to a large RMSD between the loops, but leads to a large change in dihedral angle, and, as North et al. cluster in dihedral space, the length 11 CDRs in L1-10,11,12-A are split into L1-11-1 and L1-11-2. The opposite effect can be seen for our clusters L1-11-A and L1-11-B. The central L1 loops of these two clusters (from 4IMK_C and 3MLS_M, respectively) are 1.5 Å apart, but are considered close enough in dihedral space to belong to North et al. cluster L1-11-3. Some clusters are split in North et al. due to our length-independent approach. For example, our cluster L1-13,14-A is split by length into L1-13-1 and L1-14-2 by North et al.

The smaller clusters (containing less than six unique sequences) map less well and there is usually no corresponding cluster in our work to match the cluster in North et al. One further difference between our work and that of North et al. is that North used a non-

redundant CDR set, filtering out the structures of the same antibody solved multiple times. We observed that these identical sequences can have structures with significantly different loop conformations (e.g., CDR-L1 loops with sequence TGTSSDVGGYNYVS, have been structurally characterized multiple times as part of the structures 1MCB, 1MCC, 1MCD, 1MCE, 1MCF, 1MCH, 1MCI, 1MCJ, 1MCK, 1MCL, 1MCN, 1MCQ, 1MCR, 1MCS,(Edmundson et al. 1993) and is found in conformations differing by over 1.5 Å between different PDB ids). Therefore, we made a decision to include all CDR structures, regardless of sequence redundancy. By doing so we avoid picking a structure that is non-representative due to crystal packing, or mistakes in solving the structure (Nikoloudis et al. 2014). This approach also allowed us to observe CDR sequences that can exist in two canonical states. However, it will also reduce our ability to predict conformations as an identical sequence could be found in two different structural clusters.

2.4 Conclusions

We analysed structural similarities between CDRs of different lengths and used them to generate length-independent structural clusters. Compared to the commonly used length-dependent approach, we generate a smaller number of clusters, containing more unique sequences. This improves our ability to classify CDRs into clusters by sequence alone.

Given that for a portion of CDRs the most similar available structure is one of a different length, and such structural similarity is usually matched by sequence similarity, developing CDR modelling methods that utilize this information should significantly improve prediction accuracy.

We have described how natural antibody affinity maturation processes can produce CDRs having different lengths, but similar structure. Since the probability of these processes generating insertions and deletions is relatively low, the length-independent structural similarities are likewise infrequent. Nevertheless, we believe that as new antibodies' crystal structures become available, length-variable clusters will become a more common occurrence.

We have tested our method on three large NGS datasets of CDR-L3 sequences and found that our length-independent methodology can classify ~135,000 or ~20% more sequences into clusters than standard techniques. We have also observed significant differences in distribution of CDR-L3 lengths between the structural dataset and the NGS datasets. This disparity, together with the imbalance between λ and κ chains in the structural dataset, is a major obstacle towards increasing the structural coverage of human antibody sequence space.

Since publication of the work presented in this chapter, the length-independent structural similarities have been used for general loop structure prediction in the Sphinx software (Marks et al. 2017). This demonstrates the utility of the idea outside of the canonical class context.

3 FAST CHARACTERISATION OF CDR-H3 STRUCTURE AND VH-VL ORIENTATION

3.1 Introduction

It has been estimated that a healthy human is capable of producing ~10^{11} unique antibodies (Glanville et al. 2009). Due to the size of this theoretical repertoire it is likely that at any given point in time the immune system contains at least one antibody molecule capable of weakly recognizing any protein antigen. Thus, rigorous structural characterisation of the antibody repertoire could broaden our understanding of functional properties of antibodies and aid the production of novel antibody therapeutics.

The two antibody structural features, commonly thought the hardest to model are the CDR-H3 loop and the VH-VL orientation (Teplyakov, Luo, et al. 2014). Because of this, the techniques used to model these two elements are usually computationally expensive (e.g. Marze, Lyskov and Gray, 2016; Weitzner, Kuroda, Marze, Xu and Gray, 2014) and, therefore, unsuitable for processing the large volumes of data now available.

In this chapter, we introduce two algorithms, one for the fast modelling of CDR-H3 loop and the other for VH-VL orientation prediction. The first algorithm is called Sphinx (Marks et al. 2017) and is a hybrid method combining *ab initio* and knowledge based modelling methodologies. This method was primarily developed by Claire Marks, but I improved the performance of the decoy ranking procedure within the Sphinx algorithm. The second algorithm we describe is a high-throughput VH-VL modelling procedure which I designed. The method uses interface sequence identity calculations between an input Fv sequence and a set of structural orientation templates to assign the orientation to the input sequence. We show that by expressing the identity calculation as a sparse matrix multiplication we can assign VH-VL orientations to billions of Fv sequences in tractable time.

The findings of this chapter contributed to two publications (Marks et al. 2017; Parks et al. 2017)

3.2 Methods

3.2.1 Sphinx algorithm

Sphinx is loop structure prediction software which combines ideas from knowledge-based (based on software FREAD, Choi and Deane 2010) and *ab-initio* modelling techniques. It was written by Claire Marks. The algorithm takes as input the sequence of a loop and predicts its structure (Marks et al. 2017). The flow of the Sphinx algorithm is illustrated in Figure 3.1.

In the first stage of the Sphinx pipeline, segments of protein structure, shorter than the loop to be modelled are selected from a database of loop fragments (Marks et al. 2017). These fragments are selected using two criteria. First, the sequences of the fragments

are aligned to the target loop using the Needleman-Wunsch algorithm (Needleman et al. 1970) and scored using an Environment Specific Substitution Table (ESST) (Marks et al. 2017). A number of top scoring fragments is selected and passed to the next stages in the algorithm. The number of fragments selected is calculated using the following, formula:

$$\text{number of fragments} = \begin{cases} 12.5(n-2) \text{ if } n \text{ is even} \\ 12.5(n-1) \text{ if } n \text{ is odd} \end{cases}$$

where n is the number of residues in the loop (Marks et al. 2017). The above formula was obtained empirically, on a preliminary test set of 100 loops, by measuring the relationship between RMSD of the best decoy produced by Sphinx and the number of fragments (Marks et al. 2017). The second step involves filtering the fragments by RMSD between Cα atoms of anchor residues, where the anchors are defined to be two residues before the start of the loop and two residues after the end of the loop (Marks et al. 2017). This anchor geometry criterion is equivalent to the one used in the FREAD software (Choi et al. 2010).

After the fragments have been selected, candidate loop models are built onto the anchor residues, starting from either the N-terminus or the C-terminus. Each model is built one residue at a time (Marks et al. 2017). The fragment residues that are matched in the alignment to the target residues have their ϕ/ψ angles copied directly onto the scaffold. The target residues that are not matched to any of the fragment residues have their dihedral angles sampled from residue-specific Ramachandran distributions (Marks et al. 2017).

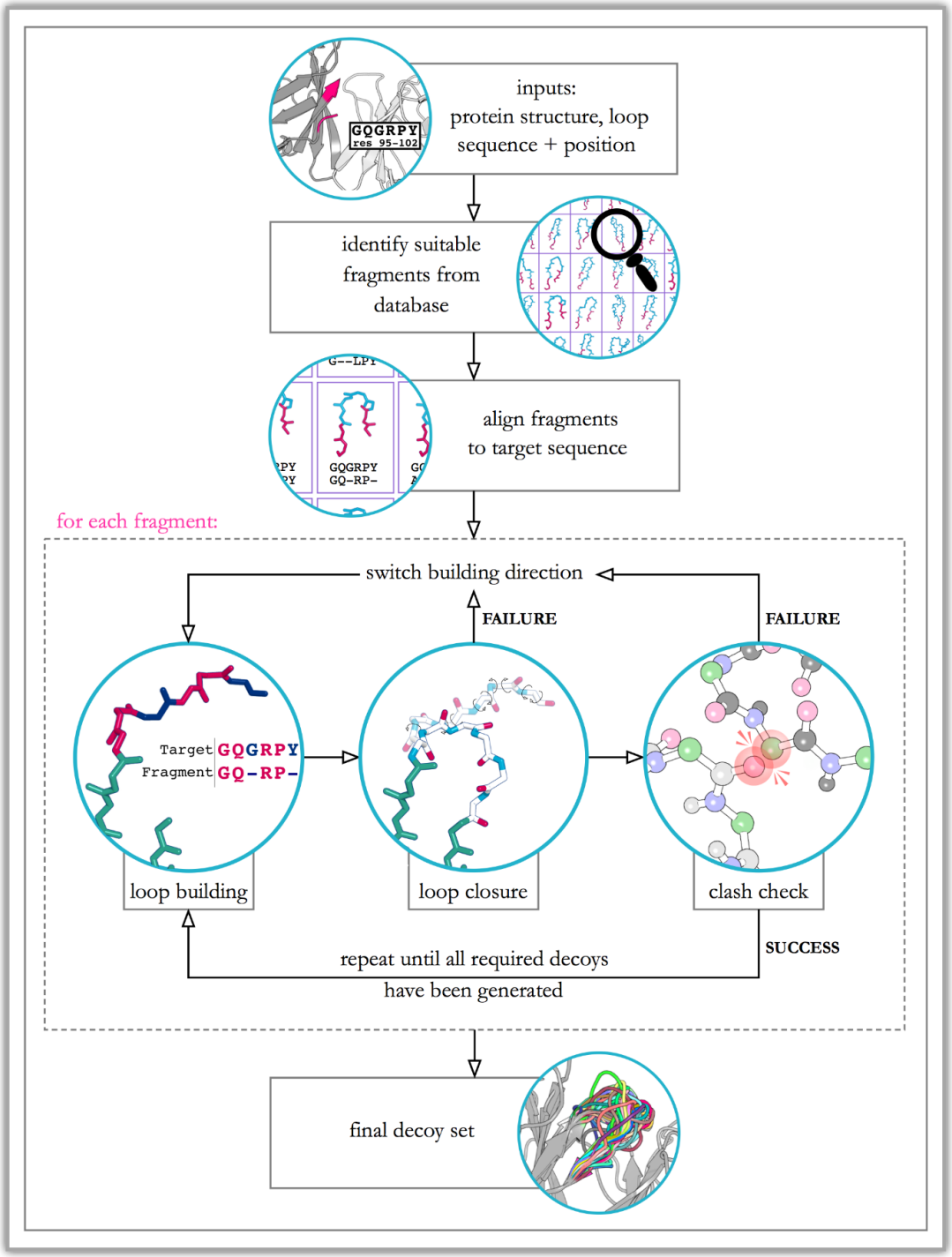

Figure 3.1: The flow of the Sphinx algorithm. The algorithm takes as input the position and sequence of the loop to be modelled. Next, it identifies suitable fragments to be used as input to an iterative decoy construction procedure. The constructed decoys are then scored using an RAPDF potential. The top scoring 500 decoys are outputted. The figure is reproduced from Marks, Deane and Shi, (2016).

After the dihedral angles for all residues have been selected, the loop model is often "disconnected", where either the C-terminus or the N-terminus (depending on the direction the loop is constructed) of the model is not bonded to the corresponding anchor on the antibody framework (Marks et al. 2017). The loop is closed using Cyclic Coordinate Descent (CCD) (Canutescu et al. 2003). If the CCD algorithm fails to close the loop within a specified number of trials, the Sphinx algorithm switches the direction from which the loop is constructed (i.e. from the N-terminus to C-terminus). This operation is known as "switching" (Marks et al. 2017).

The workflow of the Sphinx algorithm is repeated iteratively, until 100 loop models per fragment have been built, representing a possible conformational ensemble for a target loop sequence (Marks et al. 2017). The average number of output decoys per target is 11,000 (Marks et al. 2017). The pool of models is then reduced to the top scoring 500 decoys, as measured by an RAPDF potential (Samudrala et al. 1998).

To find the closest structural decoy to the true structure from these 500, a number of decoy ranking methods (Marks et al. 2017) were tested.

3.2.2 Decoy ranking methods

The problem of finding the best structural model in an initial pool of decoys is known as "decoy ranking". The top ranked decoy should be the one closest to the crystal structure. In this section, we introduce a number of decoy ranking methods and three datasets of loop structures that were used for tests.

A number of different methods have been proposed for ranking structural decoys. These scoring methods can be broadly classified into three categories: physics-based, statistics-based and consensus methods. The physics-based methods estimate the

physical energy of the structural decoy by approximating the contributions from interactions such as electrostatic, van der Walls, hydrophobic, solvation etc. Statistics-based methods use statistics derived from existing protein structures to estimate how much a given decoy resembles known structures. As such, a structural motif that is repeated frequently in the available structural data will score highly, while a previously unseen motif will score less well. Consensus methods combine the information present in many different methods.

Claire Marks tested eight methods for decoy ranking of loop models. The methods are briefly summarized in Table 3.1.

Name	**Type**	**Reference**
calRW	Knowledge-based	Zhang and Zhang 2010
calRWplus	Knowledge-based	Zhang and Zhang 2010
DFIRE	Knowledge-based	Zhou and Zhou, 2002
DOPE	Knowledge-based	Shen and Sali 2006
DOPE_HR	Knowledge-based	Shen and Sali 2006
GOAP	Knowledge-based	Zhou and Skolnick, 2011
Rosetta	Physics-based	O'Meara et al. 2015
SoapLoop	Knowledge-based	Dong et al. 2013

Table 3.1: The tested decoy ranking methods. The left column shows the name of each scoring function, the centre column shows the type of information utilized by each method and the right column shows the reference for each method.

Claire Marks created three datasets of loop structures and three corresponding fragment libraries for testing the performance of the decoy ranking methods.

The first dataset was constructed from general protein loops. First, all loop regions for all PDB entries were selected using DSSP (Kabsch et al. 1983), not including loop regions shorter than 3 amino acids or longer than 30 residues. Then, loop structures with non-identical sequences (according to PISCES (Wang et al. 2003)) were extracted from all PDB entries, keeping the structure with the highest resolution where two or more loops with the same sequence were available, forming a fragment library containing 65,108 entries. To select the test loop structures, first the loops were filtered from the library using resolution cut-off of 2.0Å and a maximum backbone atom B-factor cut-off of 30 $Å^2$. These loops were then clustered at 40% sequence identity and one loop was selected from each cluster. After this filtering step, 10 loop structures were selected at random for each length between 6 and 18 amino acids, forming a set of 70 test loop structures (full list in Appendix A4). Sphinx was used to model these 70 targets and gave a list of 500 decoys for each. All of the decoys were grafted onto the original crystal structure for scoring. This dataset was used as a "training set" to select the parameters and scoring functions for use in CDR-H3 decoy ranking.

The next two datasets contained antibody CDR-H3 loops and used a fragment library constructed by extracting all CDR-H3 loops according to the Chothia definition (Al-Lazikani et al. 1997) from the SAbDab database (Dunbar et al. 2014). This fragment library contained 3043 CDR-H3 structures, from 1848 different PDB entries, which reflected the number of structures available in SAbDab in April 2015 (Marks et al. 2017).

The second dataset was taken from Sivasubramanian et al. 2009 and contains 53 CDR-H3 structures (full list in Appendix A4). The target with the PDB id 2ai0 was removed

from this benchmark, as the CDR-H3 loop was not completely resolved. Analogous to the first dataset, after 500 decoys for each case have been produced by Sphinx they were grafted on the original antibody crystal structure for scoring (Marks et al. 2017). This benchmark was consistent with the second stage of the Antibody Modelling Assessment (AMA) (Teplyakov, Luo, et al. 2014).

The third dataset used the same 53 Fab structures as the second one. Here, Sphinx decoys are generated using models of the Fv region, rather than the crystal structure. Models of the Fv region were built using ABodyBuilder (Leem et al. 2016). This allowed us to assess the accuracy of the decoy ranking functions, when the CDR-H3 models were built and presented in a non-native environment (Marks et al. 2017). This benchmark was analogous to the first stage of the Antibody Modelling Assessment (Teplyakov, Luo, et al. 2014).

3.2.3 Machine learning for decoy ranking

The problem of constructing machine learning ranking models is known as Machine-Learned Ranking (MLR) and is used in many different areas, such as document retrieval, or sentiment analysis (Joachims 2002). A commonly used approach to MLR involves taking pairs of objects to be ranked and assigning labels to each pair describing their relative position in the ranking. In this way, the MLR problem is reduced to classification. This method is known as the pairwise approach, and transforming the ranking into a set of object pairs is known as the pairwise transform (Chen et al. 2009).

For the purpose of structural decoy ranking, we have considered pairs of decoys labelling them according to which decoy is closer to the true crystal structure, as measured by RMSD. If the first decoy was the closer one the pair would be labelled as 0 and if the second decoy was closer, the pair would be labelled as 1. This way, the decoy ranking

problem was represented as binary classification. The input features were calculated as difference between ranking scores obtained for the first decoy and the second decoy (the raw ranking scores were calculated using methods listed in Table 3.1). Since the scoring functions operated on different scales, the input features were scaled by subtracting the mean and dividing out the standard deviation. Such constructed feature-label pairs were used to train an Artificial Neural Network (ANN) ranking model.

The ANN is a supervised machine learning algorithm, inspired by the biology of a brain, which learns a non-linear function $f: X \rightarrow Y$ where X is the set of input features and Y is the set of outputs. The network is composed of an input layer, containing a number of input nodes, a set of hidden layers, transforming the input features using some non-linear activation functions, and an output layer calculating the outputs (Figure 3.2).

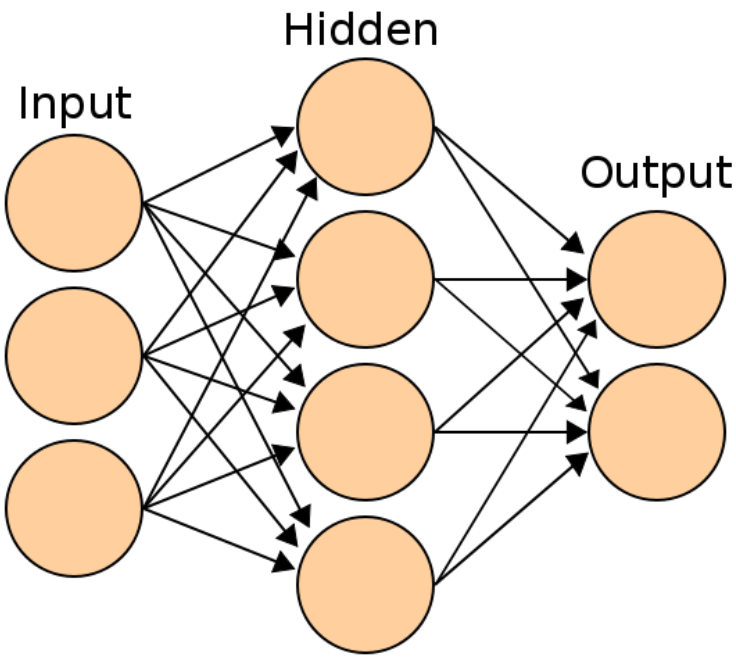

Figure 3.2: A two-layer Artificial Neural Network. The network consists of an input layer (on the left), a hidden layer (in the middle) and the output layer. The hidden layer takes as inputs the values from the input layer and transforms them using a non-linear function. The figure was

We chose to create our ranking model using the ANN algorithm because similar models have been successfully used in previous decoy ranking studies (e.g. Tan et al. 2008) and the neural network can be trained sequentially (few samples at a time). The ANN described in this chapter was constructed using the Tensorflow library (Abadi et al. 2015).

The ranking ANN architecture contained eight input nodes (one for each scoring function), one hidden layer with 12 nodes and an output layer with one binary node (see Figure 3.3 The size of the hidden layer was chosen to be close to the double of the number of input features. This choice was made as the dropout technique (Srivastava et al. 2014) was used for regularisation, which randomly occluded half of the hidden neurons during training. The model used a rectified linear unit (Hahnloser et al. 2000) activation function for the hidden layers and a sigmoid activation for the output node (to constrain the output between 0 and 1). The loss was measured using cross entropy, given by the following formula:

$$L(\boldsymbol{y},\boldsymbol{y}') = -\frac{1}{N}\sum_{n=1}^{N} y_n \log\left(y'_n\right) - (1-y_n)\log(1-y'_n)$$

Where N is the number of training examples, the vector $\mathbf{y}$ contains the true labels and the vector $\mathbf{y'}$ contains the logistic outputs of the ANN.

To optimize the parameters of the network we trained the ANN for 1,000 steps using stochastic gradient descent (Finkel et al. 2008).

The performance of our ANN ranking protocol was evaluated using 10-fold cross-validation on the first loop structure dataset, containing 70 general protein loops (see Section 3.2.1). The dataset was split into 10 subsamples containing seven targets each. Next, the neural network was trained on the first nine subsamples and tested on the remaining one. This protocol was repeated on all permutations.

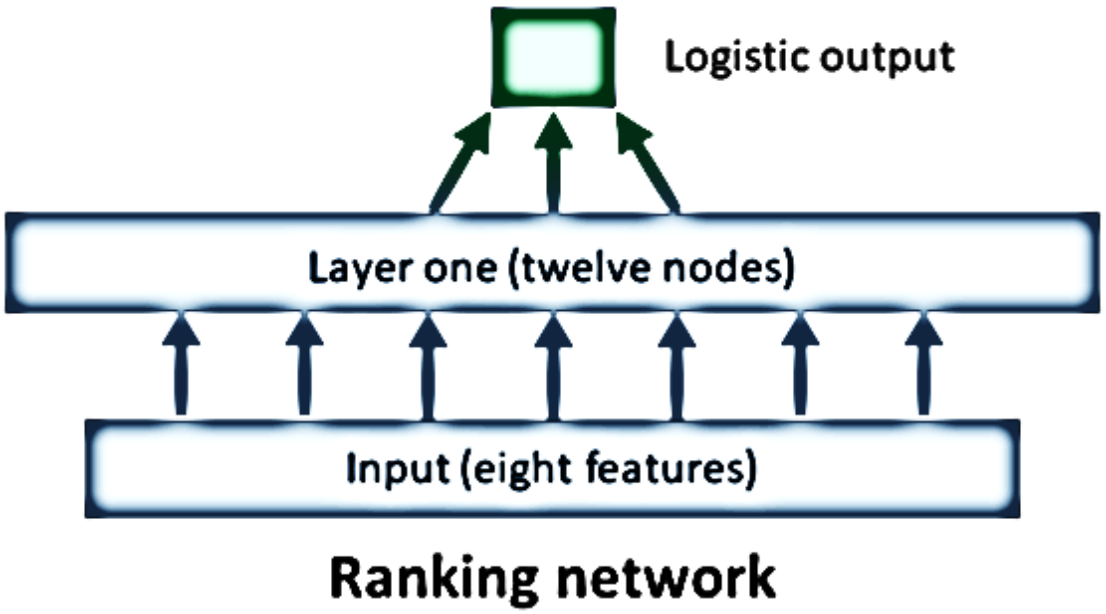

Figure 3.3: The architecture used for the ANN ranking model. First, the eight input features were transformed into a hidden layer with twelve nodes using the rectified linear unit activation function. The twelve hidden nodes were then used to calculate the binary output using a sigmoid function, which constrained the output between 0 and 1.

3.2.4 Minimization protocol

To minimize the loop models, produced by Sphinx, we used robotics-inspired loop conformation sampling protocol implemented in Rosetta, called Kinematic Closure or KIC (Stein et al. 2013).

The KIC protocol is composed of two stages – the centroid stage and the full atom stage, both of which use a Monte Carlo simulated annealing trajectory with an exponentially decaying temperature. At each step, a random segment of the loop of length between 3 and N (N is the full length of the loop) is selected and three residues (first, last and middle) are designated as pivots. Next, new values for the dihedral angles of non-pivot residues are sampled from a loop-specific Ramachandran distribution, opening the loop. To close the loop, the algorithm searches for a set of new pivot residue torsion values. The new loop shape is scored using the Rosetta Ramachandran statistical potential (Rohl et al. 2004) and the new conformation is accepted or rejected based on the Metropolis-Hastings algorithm. The centroid and full atom stages differ in how the side chains of the amino acids are represented. In the centroid stage the side chains are coarse-grained and represented using only centroids, while in the full atom stage the side chains are represented using all atoms. In the full atom stage the side chain orientations of residues within 10 Å of the loop are repacked every 20 steps through rotamer trials (Kuhlman et al. 2003).

3.2.5 Detecting crystal contacts

To detect targets with crystal contact hydrogen bonds we used the program PyMOL (Schrodinger LLC 2010). We used the maximum distance between donor and acceptor of 3.6 Å (default setting). To reconstruct the crystal lattice, we have used the function Symexp which utilized the crystal symmetry information contained within the CRYST1 record of a PDB file.

3.2.6 Quantifying the VH-VL orientation

To quantify the accuracy of orientation modelling, one must define a robust and consistent measure of VH-VL orientation. Throughout this work, we quantified the VH-

VL orientation using orientation RMSD (Narayanan et al. 2009). The benefit of this measure is that it is easy and fast to calculate. The method is described in detail in the introduction Section 1.4.4.

Having defined a VH-VL orientation measure, we turned to defining a set of interface residues, to be used in the VH-VL assignment procedures.

3.2.7 VH-VL interface residues

In order to make fast and accurate predictions of VH-VL orientations we have created a method which assigned the orientation using sequence identity of interface residues.

In order to identify the interface residues we followed a protocol inspired by the work of Dunbar et al. (2013). First, we constructed an Fv orientation template set, using antibody structures available in the SAbDab database (Dunbar et al. 2014) as of May 2016. We chose only Fv structures solved using X-ray diffraction at resolution below 3.0 Å, which resulted in an initial set of 2,578 VH-VL pairs. Next, we used ABangle (Dunbar et al. 2013) to assign six orientation angles to each Fv structure within our set. We then removed from the set structures for which any of the six ABangle metrics was more than two standard deviations from the mean. This removed any outlying Fv structures. This additional filter reduced our set to 2,460 Fv structures which we clustered at 75% sequence identity, leading to 276 distinct clusters. From each cluster, we picked the structure with highest resolution. The structures were then renumbered using the IMGT numbering scheme (Lefranc 2011). Following this, we calculated the difference in solvent accessible surface area (SASA) between separated Heavy and Light chains and the corresponding Fv structures for each of the 276 antibodies. The SASA differences were calculated using the program NACCESS (Hubbard et al. 1993). Any residue that was assigned a SASA difference greater than 0 was added to our set of

interface residues. This resulted in a set of 58 heavy chain residues and 62 light chain residues (see Appendix A5). The selected set of interface residues was used to assign VH-VL orientation using a high-throughput method described in the next section.

In order to ensure that our library of VH-VL orientation templates was unambiguous (only one orientation was assigned to any particular sequence), we created a set of non-sequence-identical antibodies. We clustered the 2,460 Fv structures at 100% interface sequence identity resulting in 989 clusters. Then, from each cluster, we picked the structure with the lowest median orientation RMSD to all other structures within the cluster (the cluster centre). This group of 989 Fv structures served as the template set which we used to model the orientation RMSD.

3.2.8 High-throughput orientation assignment

Our algorithm assigned VH-VL orientation to an input Fv sequence by calculating interface sequence identity between the Fv sequence and the set of 989 Fv interface template structures, described in the previous section. If the interface identity is over a threshold of 0.82 (see results Section 3.3.9) to any of the template structures, the orientation of the template with the highest sequence identity was assigned to the input Fv sequence.

To improve the computational performance of our sequence identity calculations we formulated the problem in the framework of sparse matrix multiplication. We encoded the 989 Fv template interface sequences using one-hot coding (Harris et al. 2013), converting each template sequence into a sparse vector 58 * 20 + 62 * 20 = 2400 positions long. The sequences were then assembled into a sparse matrix **T** of size (2400, 989). If the sequences for which we want to model the orientation are encoded in an

analogous way into a matrix **S** of size (2400, N), where N is the number of targets, then the sequence similarities can be calculated using the formula:

$$\mathbf{I} = \mathbf{S}^{T}\mathbf{T}$$

Where **I** is the (N,989) matrix containing the sequence similarities. If we divide the similarities by the number of interface residues, we obtain the required sequence identities. To perform the sparse matrix multiplication we used the cdr_matrix object from scipy.sparse library:

https://docs.scipy.org/doc/scipy-0.18.1/reference/sparse.html

By formulating our problem in this way, we took advantage of numerous sparse matrix multiplication optimizations, developed over the last decades, allowing us to process billions of potential Fv sequences.

3.2.9 Next-Generation Sequencing dataset

We demonstrated the power of our orientation assignment method on an NGS dataset containing ~5,000,000 unpaired sequences of each chain type (heavy, kappa, lambda) amounting to ~15,000,000 unpaired variable chain sequences in total (see Table 3.2).

Chain type	Number of sequences
Kappa	5,236,835
Lambda	4,134,630
Heavy	5,645,307

Table 3.2 Number of sequences of each type within the NGS dataset provided by our collaborators at UCB Pharma Ltd. The dataset contains ~5,000,000 unpaired IgM sequences of each chain type, derived from ~500 individuals.

The NGS set was collected from sequencing experiments performed by UCB pharma. The sequenced RNA samples originated from normal human spleen samples from ~500 individuals, mostly of Asian origin. The samples were sequenced using an Illumina MiSeq device at the Oxford Genomics Centre (OCG) at the Welcome Trust Centre for Human Genetics, Oxford.

The sequences were filtered for quality using the following criteria: the sequences had to be of full length (maximum two nucleotides at 5' and 3' end were allowed to be missing), and could not contain any stop codons or ambiguous nucleotide calls. Following the quality filtering, the V, D and J germlines were annotated using the IgBLAST 1.4.0 (Ye et al. 2013). If the V or the J germline could not be identified the sequence was discarded.

Each of the sequences was numbered using the ANARCI software and IMGT (Lefranc et al. 2009) scheme.

3.3 Results & Discussion

3.3.1 Overview

Sphinx is a loop structure prediction software package that combines both knowledge-based and *ab initio* modelling methods. The algorithm is capable of producing structural decoys closely mimicking the true crystal structure of the loop (Marks et al. 2017).

In this chapter, I describe how we benchmarked different methodologies for finding these close structural models. We also investigated the impact of structural energy minimization and of crystal environment on the ranking performance. The ranking results are demonstrated through the commonly-used metrics of the top decoy RMSD and the best-out-of-five decoy RMSD. The top decoy RMSD measured the structural similarity of the top ranked model to the crystal structure of the loop, while the best-out-of-five RMSD measured the structural similarity of the closest model in the top five ranked decoys. The RMSD values were calculated by first superimposing protein structures, without the loops to be compared, and then calculating the deviation of three-dimensional coordinates of backbone atoms that constitute the loop models.

In the second part of the results, we describe a high-throughput method for VH-VL orientation assignment. The algorithm modelled the orientation by comparing the VH-VL interface residues between the sequence to be modelled and a set of structurally characterised orientation templates. By expressing the sequence identity calculations through sparse matrix multiplication, we enabled our method to process billions of Fv sequences in tractable time.

3.3.2 Decoy ranking methods performance

Claire Marks tested the performance of each of the ranking methods, described in Table 3.1, on the first benchmark dataset, containing general protein loops (see methods

Section 3.2.2 and Appendix A4). The results are shown in Figure 3.4. A good structural model of a loop would typically have backbone RMSD below ~1.5 Å. The figure shows that none of the methods were capable of finding an acceptable model most of the time.

As the tested decoy ranking methods were not able to recall reliably the best model out of a pool of structural decoys produced by Sphinx, we constructed a machine learning decoy ranking method using Artificial Neural Network (ANN) ranking model (see methods Section 3.2.3).

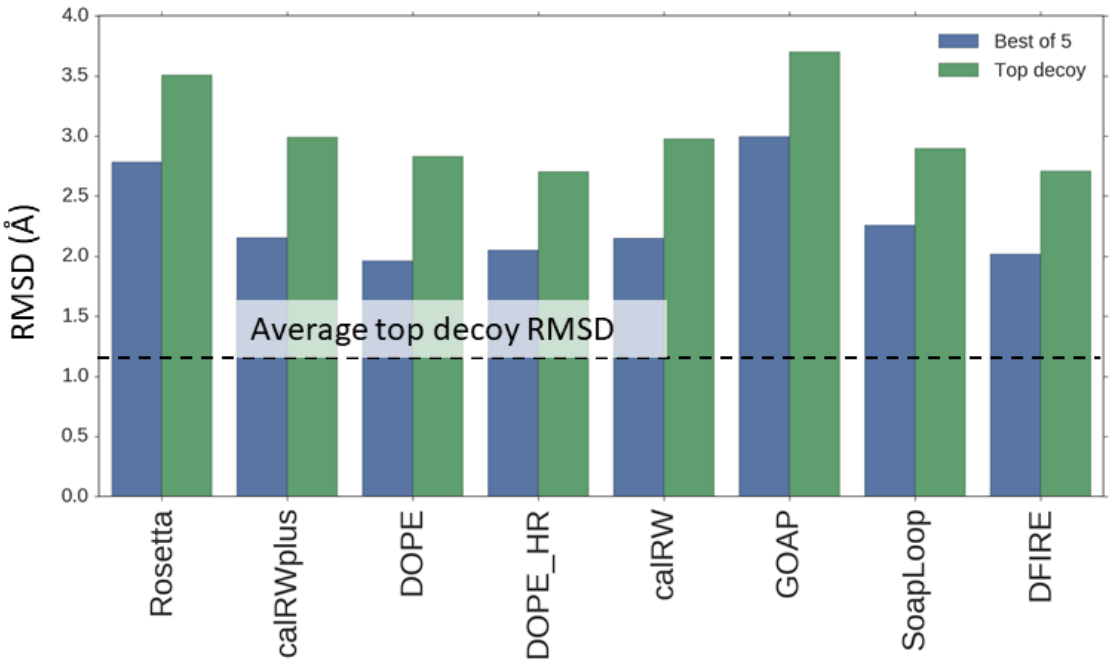

Figure 3.4: The results of ranking the general protein loop models by each method tested. The dashed line shows the average RMSD of the best decoy in the set produced by Sphinx. The average RMSD of the top scoring decoy, as found by each method, is shown using green bars. The average best RMSD in the top 5 decoys, as found by each method, is shown using blue bars. This image shows that even though Sphinx is capable of producing loop models resembling the crystal structure, the ranking methods are not capable of reliably finding that best model. The figure was created using data calculated by Claire Marks (Marks et al. 2016).

3.3.3 Consensus methodology performance

In this section, we describe performance of our ANN ranking function. We trained the neural network to combine the scores of the eight ranking methods (see Section 3.2.2), and tested its performance. We have tested our ranking model using a 10-fold cross-validation strategy on the general protein loop dataset (see Figure 3.5). The ranking models performance was similar to that of the other methods and failed to improve the overall ranking results.

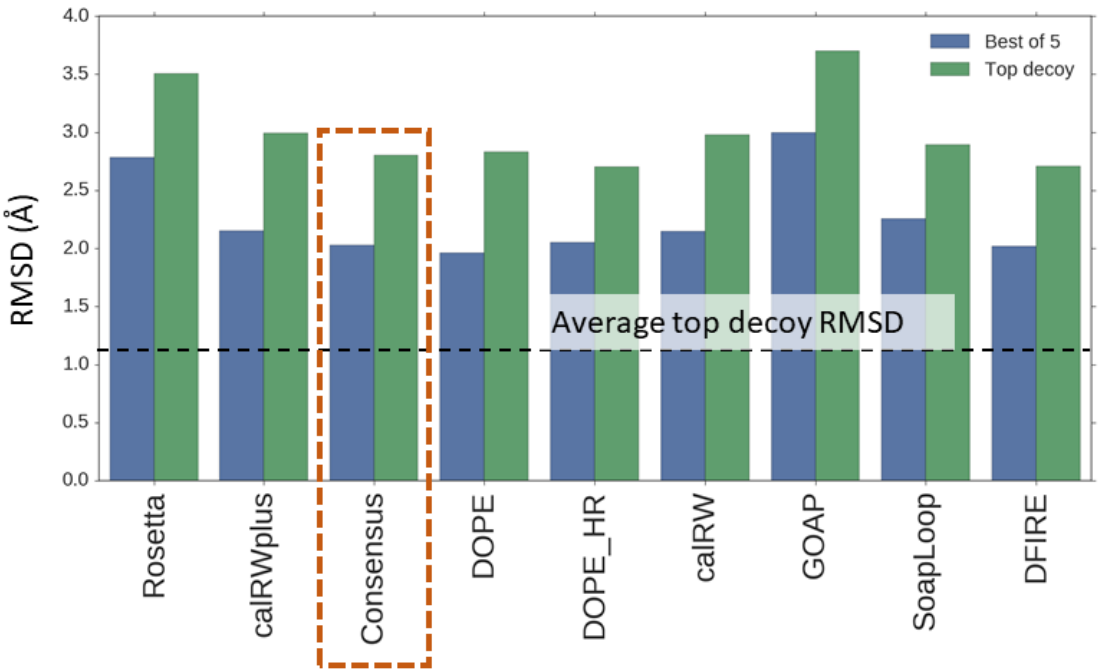

Figure 3.5: The results of ranking the general protein loop models by the consensus method. The dashed line shows the average RMSD of the best decoy in the set produced by Sphinx. The average RMSD of the top scoring decoy, as found by each method, is shown using green bars. The average best RMSD in the top 5 decoys, as found by each method, is shown using blue bars. The results obtained using the consensus method are indicated by a dashed orange rectangle. The image shows that the performance of the consensus method is similar to other decoy ranking methods.

We have investigated possible reasons behind the failure of our ranking function. Figure 3.6 shows the Pearson correlation coefficients between all the methods studied and the RMSD for two example targets. The figure shows that when an accurate model can be found by one of the methods, the other methods also perform well; this is indicated by high correlation between the individual scores and the RMSD. In contrast when the loop models are difficult to rank, all methods perform poorly, which is indicated by low correlation between the scores and the RMSD. This suggests that when the loop target is relatively easy to model all ranking methods convey similar information. In contrast,

when the target is difficult to predict, all methods perform poorly, contributing little to the consensus ranking function.

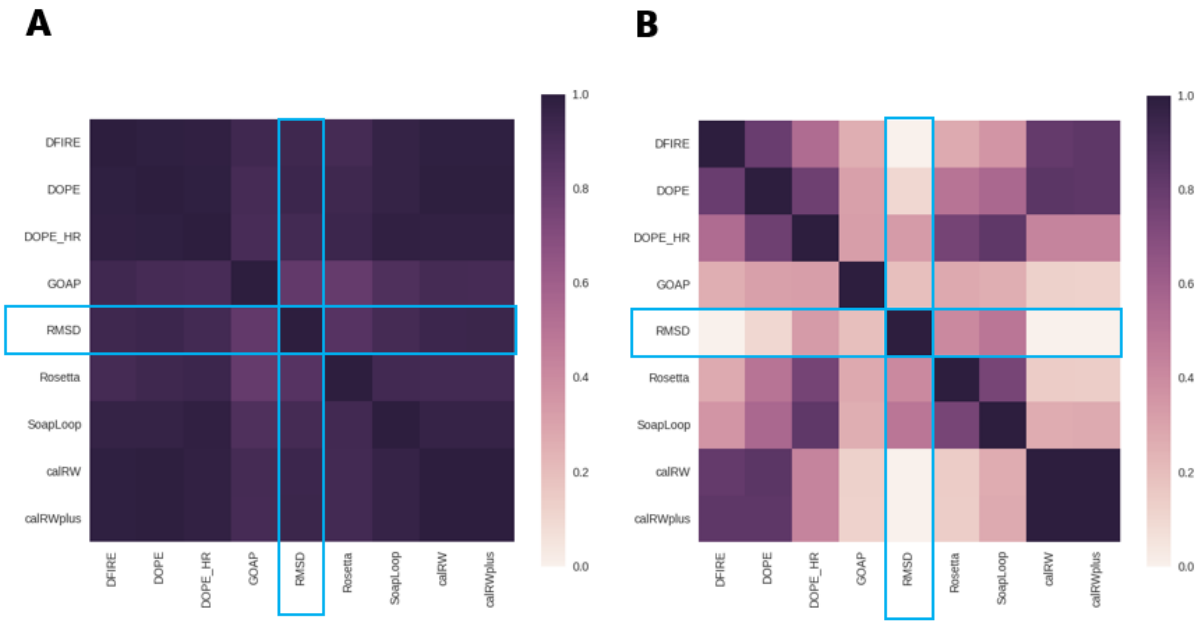

Figure 3.6: The Pearson correlation matrices. The heatmaps show the correlations between different ranking scores and the RMSD (highlighted in blue). The shade of purple indicates the strength of the correlation, with 0 indicating no correlation and 1 indicating perfect correlation. Panel A visualizes the ranking results for a loop from PDB ID 3gve. The models for this loop were ranked correctly by all methods which is indicated by high correlation between the scores and the RMSD. Panel B shows the ranking results for a loop from PDB id 1n08. None of the investigated methods could rank the models of this loop correctly which is evidenced by the low correlation coefficients.

In the next section, we consider possible reasons behind the observed results and attempt to improve the ranking performance by computationally relaxing loop decoys.

3.3.4 Decoy energy minimization

We found that many of the decoys produced by Sphinx contain steric clashes that would not occur in a true protein structure (see Figure 3.7). We observed that the ranking

methods could not automatically dismiss these clashing decoys, instead assigning them scores uncorrelated with the RMSD. In this section, we describe how we applied energy minimization protocols, implemented in the Rosetta software, to relax the decoy structures. We show that employing this minimization strategy improved the performance of the decoy ranking methods.

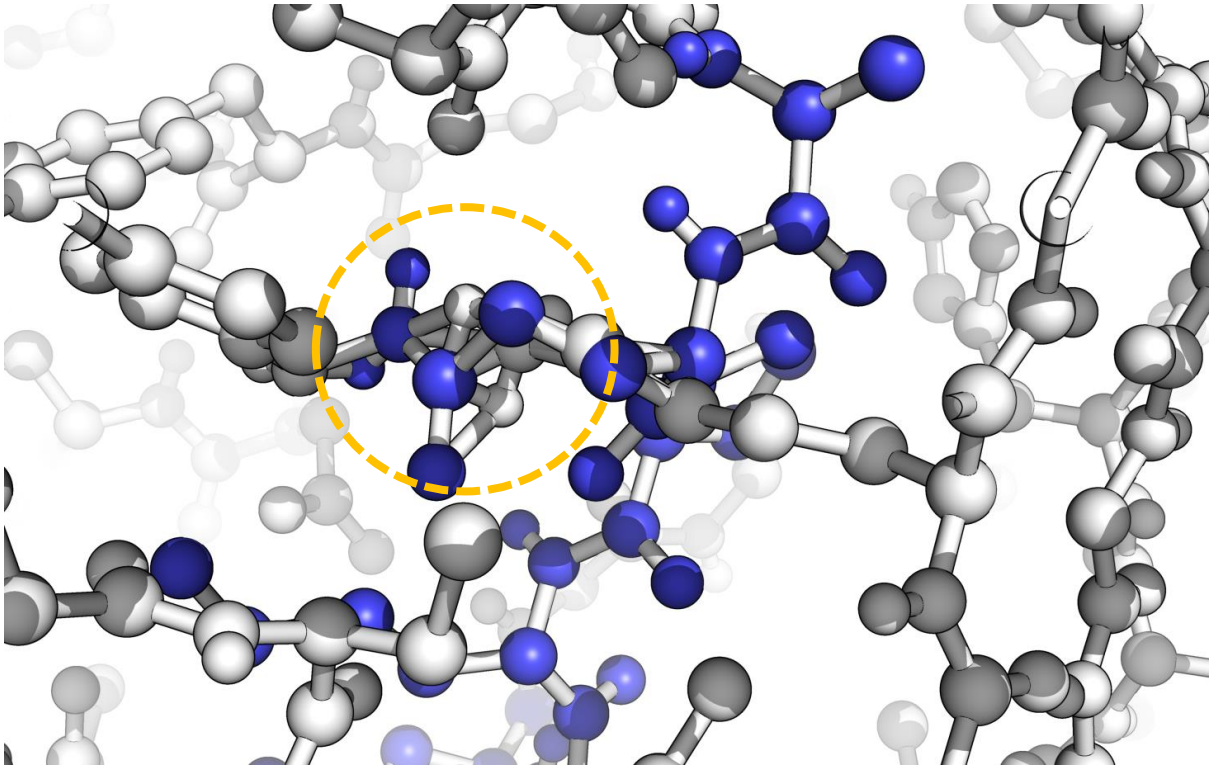

Figure 3.7: A structural clash example. The loop model is shown in blue while the protein framework is shown in white. A side chain of a Glutamine (in blue) is clashing with side chain of Lysine (in white). The clash is indicated by a dashed yellow circle. Such clashes would create unreasonable scores and hinder the performance of decoy ranking methods.

We tested the influence of model minimization on decoy ranking performance using our general protein loop dataset (see methods Section 3.2.2). We subjected the initial loop decoys to short runs of the Rosetta's KIC protocol (see methods Section 3.2.4). Each decoy was minimized using three centroid cycles and three full atom cycles. Such short

minimization took about 60 CPU seconds per decoy and about 8 CPU hours per target (500 decoys).

The impact of model minimization on ranking is shown in Figure 3.8. The minimization improved the ranking of loop models. The best performing method, SoapLoop, when run on minimized structures gave an RMSD of 1.50 Å, compared to 2.1 Å before minimization. The improvement in performance was also indicated by higher Pearson correlations between the scores and the RMSD, for the DOPE_HR, GOAP, Rosetta and SoapLoop methods (see Figure 3.8). The higher correlation scores indicated that the ranking was improved throughout the entire decoy set.

It is important to note that the minimization protocol did not, in general, improve the RMSD of the individual decoys (Figure 3.9). Instead, by relaxing the loop backbone and side chain orientation, it shifted the conformations to more energetically favourable ones as predicted by Rosetta scoring function, improving the ranking.

Despite the improvement of the overall ranking results, for some targets the minimization step decreased the ranking accuracy. In the next section, we describe our investigation of the impact of crystal hydrogen bonds on performance of ranking methods.

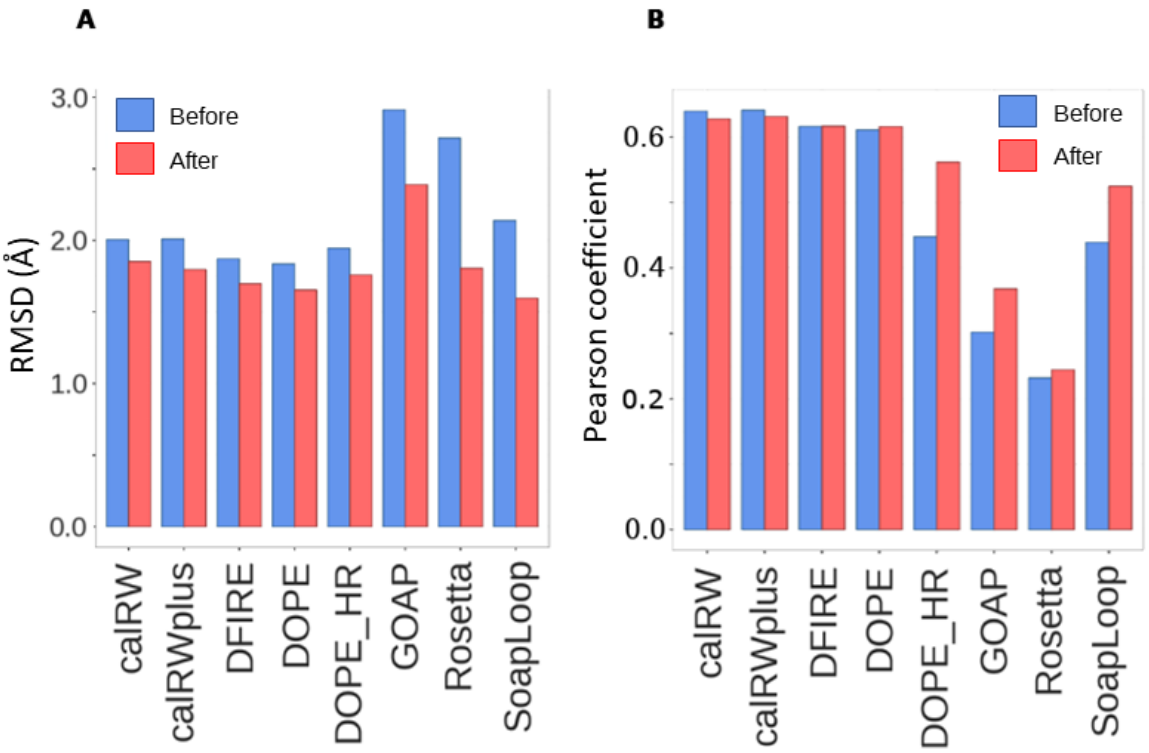

Figure 3.8 The effect of model minimization on decoy ranking performance. Panel A shows the improvement in the Best-out-of-5 metric. The blue bars show the RMSD of the best-out-of-five decoy before the minimization while the red bars show the RMSD after the minimization. Panel B shows the change in average Pearson correlation between scores produced by individual methods and the RMSD. The blue bars show the correlation before the minimization and the red bars show the correlation after minimization. Overall, the figure shows that the minimization improved the ranking performance, as indicated by lower RMSD values and higher correlation coefficients.

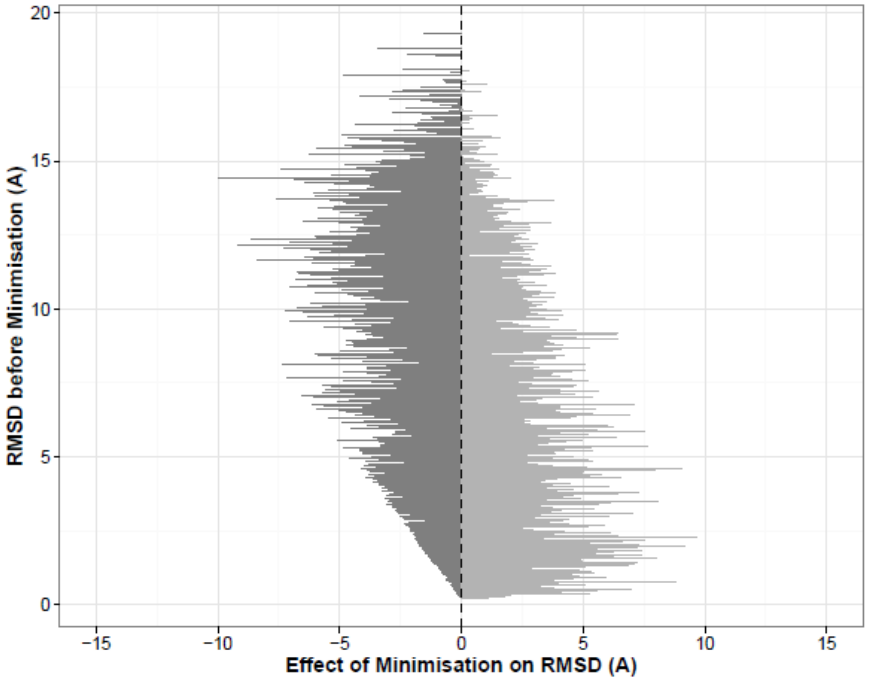

Figure 3.9: The effect of KIC minimization on individual decoy RMSD. This figure shows the change in distance of individual loop models to the native conformation, for our general protein loop dataset. The y axis shows the original RMSD distance between each model and the native conformation, while the x axis shows the change in decoy RMSD after minimization. The figure shows that, in general, the minimization did not improve individual RMSD scores, often pushing the loop conformation further from the native state. The figure was reproduced from Marks et al., (2017).

3.3.5 Crystal contact hydrogen bonds

We observed that many of the loop modelling targets, for which the KIC minimization failed to improve the ranking results, contained interactions with residues in neighbouring crystal units. As we were modelling only a single copy of the protein structure, not the complete crystal, we were unable to capture the impact of such interactions on the loop conformation. When crystal contact hydrogen bonds exist, they may have an influence on the structure of the loop (Figure 3.10). Such crystal contact hydrogen bonds could stabilize loop conformations that would be less favourable outside

of the crystal context, confusing the decoy scoring functions and the minimization protocol.

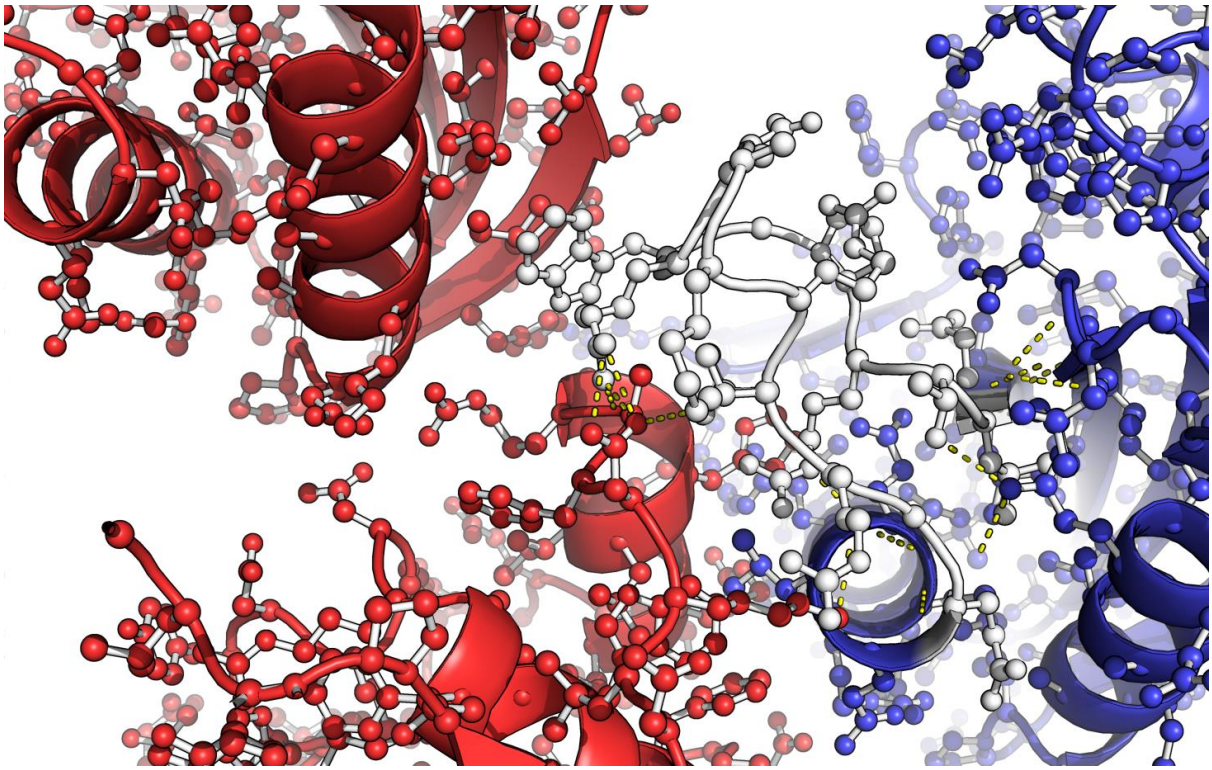

Figure 3.10 The influence of crystal contact hydrogen bonds on the loop structure. The figure shows one of the loop targets from our general protein loop dataset (PDB ID 3lill, chain A, residues 81-94). The loop to be modelled is shown in white, while the rest of the protein is shown in blue. The neighbouring crystal units are shown in red. The hydrogen bonds are shown in yellow. The loop contained many hydrogen bonds to the residues on the neighbouring crystal units, potentially influencing the conformation.

The ranking results obtained by removing the targets with crystal contact hydrogen bonds are shown in Figure 3.11. The best-out-of-five RMSD on this set using the SoapLoop scoring function was 1.05 Å, down from 1.50 Å. This improvement arose in part because longer loops were more likely to contain crystal contact hydrogen bonds and were also harder to model, but looking at the results on a per-length basis we saw

that the improvement in ranking performance was largely length-independent (Figure 3.11 B).

By minimizing the loop models and removing targets with crystal hydrogen bonds we obtained the best-out-of-five decoy RMSD of 1.05 Å, on our general protein loop dataset. This result indicated that most of the time we could find a loop model which was very close to the native conformation. In the next section, we describe how we applied our protocol to the antibody CDR-H3 benchmarks and discuss the results.

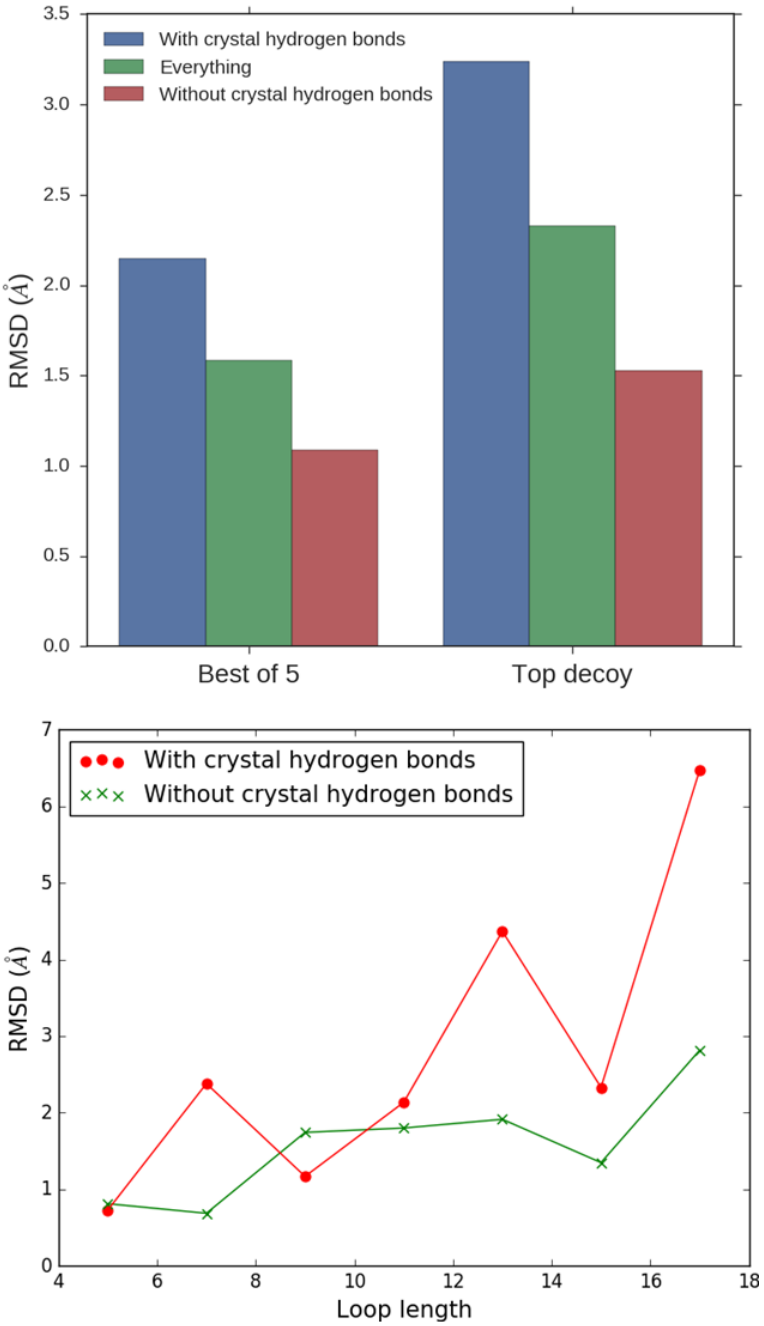

Figure 3.11 The effect of removing targets with crystal contact hydrogen bonds on ranking performance. The top panel shows the comparison of the ranking performance of the SoapLoop method on the set of targets containing crystal contact hydrogen bonds to the set of targets without the crystal contact hydrogen bonds. The mean RMSD of decoys created for targets containing crystal hydrogen bonds is shown using blue bars, the mean RMSD of decoys created for targets which do not contain crystal hydrogen bonds is shown in red, while the mean RMSD

of all predictions is shown in green. The bottom panel shows the RMSD of top ranked decoy on a per-length basis. The top decoy RMSD of targets containing crystal hydrogen bonds is shown in red and the top decoy RMSD of the targets which do not contain crystal hydrogen bonds is shown in green. The figure shows that removing targets with crystal contact hydrogen bonds improved the ranking results and this effect was largely length-independent.

3.3.6 Antibody benchmarks

In this section, we describe how we tested the performance of Sphinx algorithm on the antibody CDR-H3 datasets (see Section 3.2.2). Our results show that by using the techniques described above, we were capable of reliably modelling the CDR-H3 loop to a high accuracy. We also compared the performance of Sphinx to RosettaAntibody, one of the leading CDR-H3 modelling protocols. We show that Sphinx can achieve a better modelling accuracy, at a significantly reduced runtime.

Following our previous results, before testing Sphinx, we have removed 15 loops with crystal contact hydrogen bonds from our CDR-H3 datasets. This left us with 39 targets to test the performance of the algorithm. Claire Marks modelled the CDR-H3 loops on the native crystal frameworks (dataset 2, methods Section 3.2.2) and on the modelled frameworks created by the program ABodyBuilder (Leem et al. 2016) (dataset 3, methods Section 3.2.2). Five hundred decoys were created per target for each dataset. Comparing the performance of the algorithm on these two datasets allowed us to quantify how much loop prediction accuracy was affected by overall structure prediction accuracy.

A

	Crystal Structures (no minimisation)			Crystal Structures (with minimisation)		
	Best	Top	Top 5	Best	Top	Top 5
Median	1.05	1.74	1.5	0.71	2.31	1.05
Mean	1.18	2.43	1.93	0.94	2.5	1.52
No<1.0A	19	13	13	27	9	17
No<2.0A	34	21	24	37	17	30
No<3.0A	38	24	30	38	25	34

B

	Model Structures (no minimisation)			Model Structures (with minimisation)		
	Best	Top	Top 5	Best	Top	Top 5
Median	1.51	2.18	1.79	1.18	3.12	2.53
Mean	1.73	3.21	2.58	1.41	3.26	2.6
No<1.0A	10	2	5	11	4	5
No<2.0A	28	18	22	34	10	14
No<3.0A	35	24	27	37	18	27

Figure 3.12 The performance of Sphinx on the CDR-H3 benchmarks. The figure shows the number of targets that could be predicted at accuracy higher than specified thresholds (1.0, 2.0, 3.0 Å). Panel A shows the performance of CDR-H3 modelling on native crystal structure of the framework (dataset 2) while panel B shows the performance on the modelled framework (dataset 3). The minimization step improved the results for best-out-of-five metric on the crystal structures, but reduced the performance on model structures. The table showed that Sphinx can produce accurate models of CDR-H3 loop conformation when the crystal structure of the framework is known. The figure was reproduced from Marks et al., (2017).

The results are summarized in Figure 3.12. As anticipated, the algorithm performed better when the crystal structure of the framework was known. The minimization improved the best-out-of-five RMSD scores for CDR-H3 models grafted on the crystal structure of the framework, but made the scoring more difficult for CDR-H3 models grafted on a model of the framework. This result underlined the importance of the environment of the loop. Overall the results showed that when using Sphinx for CDR-H3 modelling we can expect to obtain an accurate model about ~50% of the time when the crystal structure is known and about 25% of the time when the structure of the

framework is not available. The performance is better in comparison to other commonly-used software for CDR-H3 prediction, as described in the next section.

3.3.7 Comparison to RosettaAntibody

We compared the modelling performance of Sphinx to that of the RosettaAntibody software (Weitzner et al. 2014). Using the RosettaAntibody software we created 500 decoys for each of the 39 targets supplying the algorithm the native crystal structure of the framework and applying the recommended options suggested on the Rosetta website (https://www.rosettacommons.org/docs/latest/application_documentation/antibody/antibody-model-CDR-H3). Producing each decoy with RosettaAntibody took about an hour of CPU time, adding up to 500 hours of CPU time per target. This is in sharp contrast to Sphinx which, on average, could output a complete prediction in 14 hours of CPU time. The results are shown in Figure 3.13. Sphinx and RosettaAntibody were equally capable of producing a robust set of initial decoys, as indicated by the "Best" column of Figure 3.13. After ranking the decoys, Sphinx outclassed RosettaAntibody achieving average best-out-of-five RMSD of 1.52 Å (compared with 2.35 Å for Rosetta) and predicting 30 targets at best-out-of-five RMSD below 2.0 Å (compared with 20 targets for RosettaAntibody).

The results showed that Sphinx outperforms the state-of-the-art CDR-H3 loop modelling software and uses much less computational resources.

Having described the improvements to the CDR-H3 prediction Sphinx software, in the following sections we focused on our high-throughput VH-VL orientation prediction protocol.

A Sphinx (with minimisation)

	Best	Top	Top 5
Median	0.71	2.31	1.05
Mean	0.94	2.5	1.52
No<1.0A	27	9	17
No<2.0A	37	17	30
No<3.0A	38	25	34

B Rosetta

	Best	Top	Top 5
Median	0.75	3.25	1.48
Mean	1.07	3.38	2.35
No<1.0A	25	8	17
No<2.0A	35	13	20
No<3.0A	37	18	26

Figure 3.13: The comparison of modelling performance of Sphinx and of RosettaAntibody. The figure showed the number of targets that could be predicted at accuracy higher than specified thresholds (1.0, 2.0, 3.0 Å). Panel A summarizes the performance of Sphinx while panel B shows the performance of RosettaAntibody. The models produced by Sphinx were more accurate and required less computational resources. The figure was reproduced from Marks et al., (2017).

3.3.8 VH-VL orientation flexibility

The orientation between the Variable Heavy (VH) and Variable Light (VL) domains of the Fv region of the antibody remains one of the most challenging features of antibody structure to model accurately (Dunbar et al. 2013). It has been shown that the VH-VL orientation is subject to rotational flexibility and that the antibody binding events sometimes constitute induced-fit behaviour (Dunbar et al. 2013). Because of this we need to define an "identicality" orientation RMSD threshold, below which the Fv orientations can be treated as equivalent. Towards that purpose we measured the orientation RMSD (see methods Section 3.2.6) between sequence identical antibody pairs.

To find the sequence identical antibody pairs we used a set of 2,460 Fv structures selected from the SAbDab database (see methods Section 3.2.7). We found 7,552 VH-VL pairs with 100% amino acid identity over the whole Fv region. We calculated the orientation RMSD distances between these identical pairs. The results are shown in

Figure 3.14. The average orientation RMSD between identical pairs was 0.6 Å with standard deviation of 0.5 Å. We chose our identicality threshold to be 1.5 Å, which explained 95% orientation RMSD values between sequence identical VH-VL pairs.

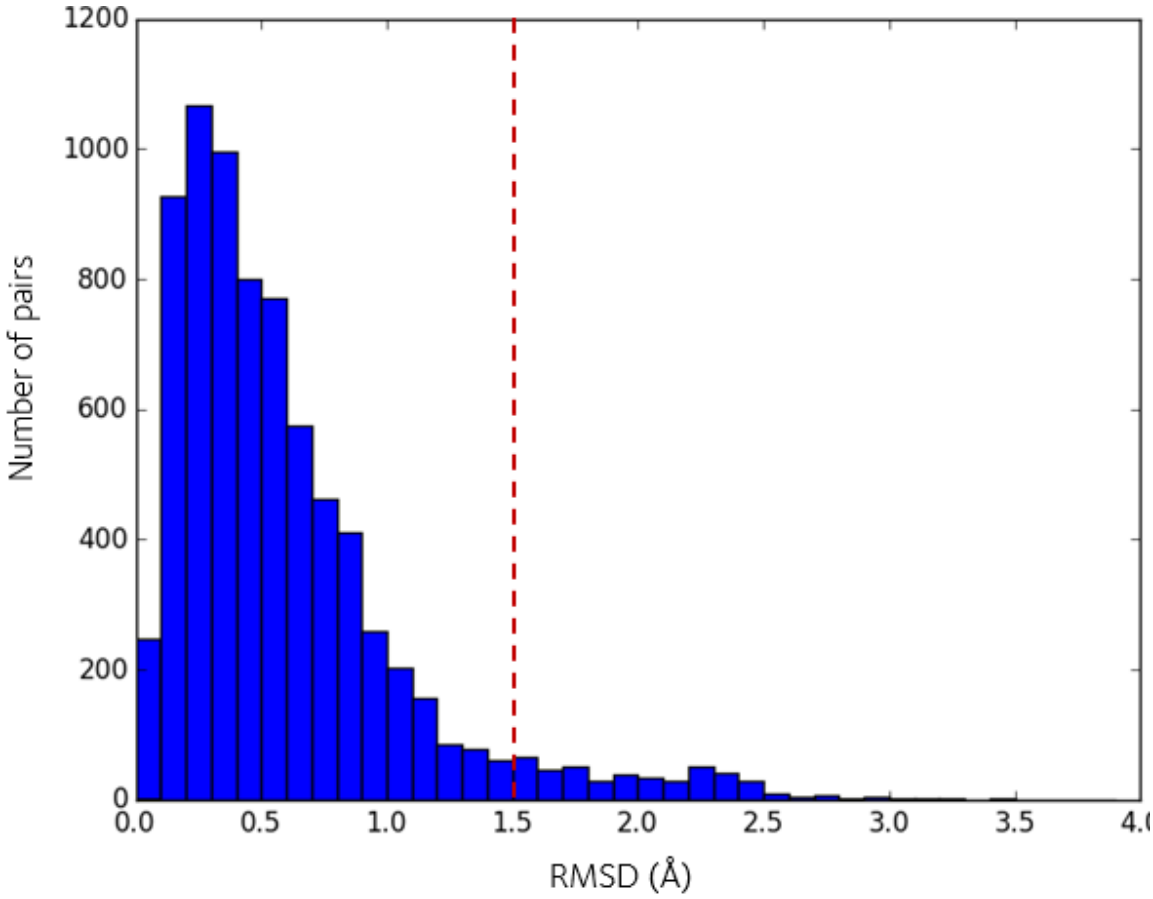

Figure 3.14: The orientation RMSD distribution for 7,552 sequence identical X-ray structure pairs. The histogram shows the number of pairs contained within each orientation RMSD bin. Most RMSD values (95%) lie below 1.5 Å (indicated by a dashed, red line) which we chose as our identicality threshold.

The SAbDab database contained one solution NMR structure of a human single-chain Fv fragment (scFv) (PDB id 2kh2). This presented an opportunity to compare our analysis of X-ray diffraction structures to the antibody NMR data. We have extracted individual

models from the PDB id 2kh2 (77 models) and calculated the orientation RMSD between the model pairs (2,926 pairs). The results are shown in Figure 3.15.

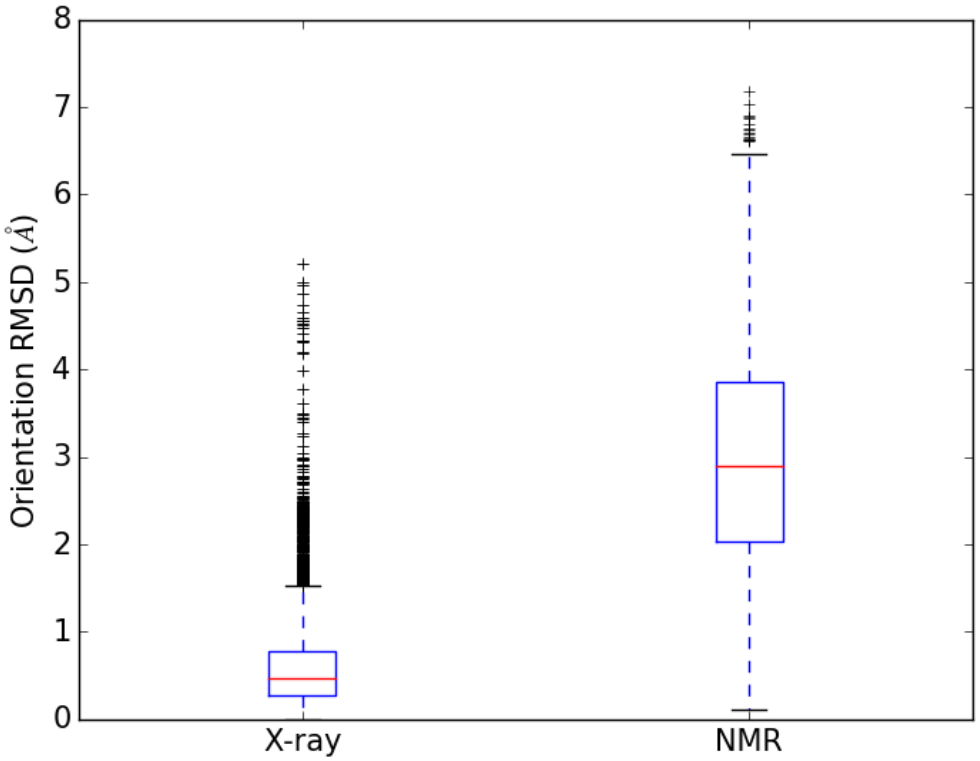

Figure 3.15 The orientation RMSD comparison between sequence identical X-ray structures and the solution NMR structure with PDB id 2kh2. The X-ray box shows the range of orientation RMSDs calculated for 7,552 sequence identical X-ray structure pairs. The NMR box shows the range of orientation RMSDs calculated for 2,926 pairs of solution NMR structure pairs. The spread of orientation RMSDs for the NMR structures was much larger than for the X-ray crystallography structures, with RMSD values reaching over 7 Å.

The NMR data suggested that the flexibility of the VH-VL interface could be much larger than suggested by the X-ray results.

In the following sections, we use the more conservative value of 1.5 Å as the threshold of identicality. This put more confidence in our results; nevertheless, it is important to note that the VH-VL orientation may be more flexible than suggested by the X-ray data.

3.3.9 VH-VL orientation prediction

Having defined the orientation RMSD threshold below which the VH-VL orientations could be considered equivalent, we created a high-throughput method for assigning the orientation, based on interface sequence identity calculations.

First, we identified a set of amino acids which have been observed to lie at the Heavy-Light chain interface. We calculated the SASA values for all sequence positions, defined using IMGT numbering (Lefranc et al. 2009), found in a set of 2,460 structurally characterised antibodies (see methods Section 3.2.7), for all Heavy and Light chain structures separately and for complete Fv structures. Next, difference is SASA value (dSASA) between the two configurations was calculated for each residue. Any residue with dSASA greater than 0.0Å^2 was classified as belonging to the VH-VL interface. We identified 58 heavy chain interface residues and 62 light chain interface residues (see Appendix A5).

Next, we searched for an interface sequence identity threshold, above which most of the structure pairs have equivalent orientations (orientation RMSD below 1.5 Å). We calculated interface sequence identities for all Fv structure pairs, within our set of 2,460 structurally characterised antibodies (see Figure 3.16). We observed that at sequence identities above ~0.82 about 80% orientation RMSD values were below the threshold of 1.5 Å (see Figure 3.16). Subsequently, we chose the 0.82 sequence identity cut off as a condition for assigning the VH-VL orientation to the Fv sequence to be modelled. If the interface sequence identity of a given Fv sequence were below 0.82 to all structurally characterised antibodies, we would classify the sequence as un-modellable in terms of VH-VL orientation.

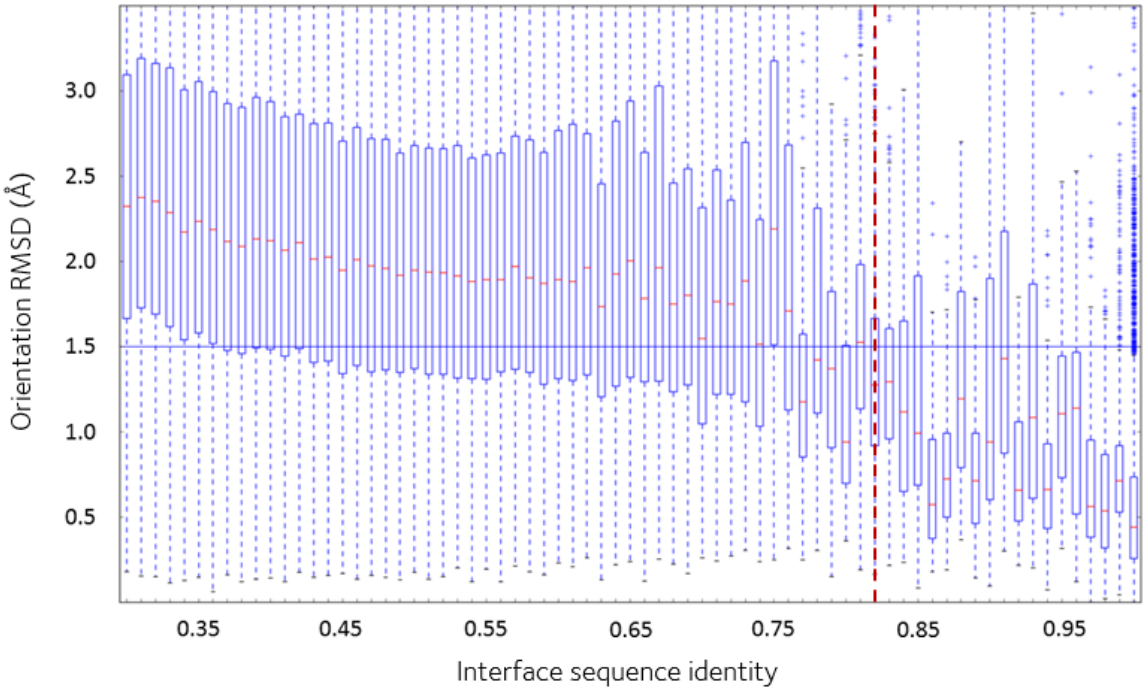

Figure 3.16 The orientation RMSD distributions at different interface sequence identity. The orientation RMSD identicality threshold of 1.5 Å is indicated by a blue horizontal line. At a sequence identity threshold of 0.82 (indicated by a vertical red line) about 80% of Fv pairs were below the threshold of 1.5 Å.

As our algorithm assigned the orientation using sequence identity calculations, the assignment results would be ambiguous if multiple templates had the same interface sequence. In order to build an unambiguous VH-VL orientation library, we filtered out sequence identical antibodies from our set, clustering the set of 2,460 Fv structures at 100% interface sequence identity and selecting from each cluster the Fv with lowest median orientation RMSD to other cluster members. This reduced the number of Fv orientation templates from 2,460 to 989 (see methods Section 3.2.7).

To improve the performance of our method, we expressed the sequence identity calculations as sparse matrix multiplication (see methods Section 3.2.8). This allowed us to perform billions of comparisons in tractable time, as demonstrated in the next section

where we show how we benchmarked our method on an NGS dataset of antibody sequences.

3.3.10 NGS dataset benchmark

To test the speed of our method we predicted the VH-VL orientation for sequences in our NGS dataset (described in methods Section 3.2.9).

We clustered the heavy and lambda sequences at 90% sequence identity using the program CDHIT (Fu et al. 2012) with default settings, choosing the representative sequence (as defined by CD-HIT) from each cluster. This reduced our dataset to ~50,000 heavy sequences and ~100,000 lambda sequences.

Since the sequences in our dataset were unpaired, they had to be coupled before assigning the orientation. The research on mechanisms responsible for VH-VL pairing is inconclusive, but the recent work suggests that most of the germline genes pair at random, with only a few having any pairing preferences (e.g. Brezinschek, Foster, Dörner, Brezinschek and Lipsky, 1998; DeKosky et al., 2014; Jayaram, Bhowmick and Martin, 2012). One of the benefits of using the interface residue identity for VH-VL orientation prediction is that all VH-VL pairs for which we can assign the orientation will have interfaces resembling those contained within the existing antibody structural data. Therefore, we could subvert the pairing problem by pairing all light sequences with all heavy sequences and then retaining only the sequence pairs above the 0.82 interface sequence identity threshold.

We paired all the lambda sequences to all the heavy sequences resulting in ~5,000,000,000 potential Fv sequences. We encoded this set into a sparse matrix and calculated the sequence identities as outlined in methods Section 3.2.8. We found that

at 0.82 interface sequence identity cut-off we were able to model the orientation of ~6,000,000 Fv sequences, or about one sequence in a thousand.

The entire analysis required ~5 x 10^{12} sequence identity calculations. The time required for this computation was about ~4 days on 40 CPUs, which was tractable. In Chapter 5 we describe how we used the VH-VL orientation assignment protocol in our computational antibody design pipeline.

3.4 Conclusions

In this chapter, we discussed decoy ranking for the Sphinx loop modelling algorithm. We showed that commonly used decoy scoring methods performed poorly when used for loop model ranking. Combining the information from different scoring functions into a consensus method did not improve the ranking results.

Minimizing the loop models, prior to ranking, improved the results by reducing the amount of conformational strains and unfavourable interactions between side chains. We also observed that if there were crystal contact hydrogen bonds between the loop and the neighbouring crystal units, the minimization tended to move the loop away from the conformation seen in the crystal. We therefore removed loops from our dataset that contained crystal contact hydrogen bonds; this improved our ranking results.

Finally, we showed that Sphinx outcompeted the state-of-the-art CDR-H3 modelling software at a reduced computational cost.

In the next segment of the results section, we discussed the problem of VH-VL orientation prediction from the point of view of high-throughput modelling of antibody NGS data. We analysed the VH-VL orientation flexibility using the antibody structures available in SAbDab and defined an orientation RMSD threshold, below which the

orientations could be considered identical. Through expressing the sequence identity calculations as sparse matrix multiplication, we could assign the VH-VL orientation to large volume of sequences in a tractable time.

The Sphinx algorithm and the VH-VL orientation modelling method tackled the prediction of two antibody structural features that are widely considered most difficult to model. The low computational resource cost of these methods makes them ideal for processing large volumes of antibody NGS data, created by modern sequencing methods.

4 NOVEL ANTIBODY STRUCTURAL FEATURE DETECTION

4.1 Introduction

A biological sequence motif is a pattern of amino acids that is thought to have a functional significance (Bork et al. 1996). These patterns can be formed by residue-residue interactions, such as hydrogen bonds, salt bridges, entropic forces etc. In Chapter 2 we mentioned that antibody Complementarity-Determining Regions (CDRs) form a relatively small number of structural conformations (except CDR-H3) and that those conformations are related to the underlying sequence patterns. In this chapter, we design a methodology for finding such sequence patterns in Next-Generation Sequencing (NGS) datasets of antibody CDRs. The method could be used to identify novel canonical classes and to correlate NGS sequences with the previously observed ones.

The algorithm uses autoencoder (Hinton et al. 1994) neural-networks to compress the input sequences to a low-dimensional space. The compressed features, calculated for

the input sequence dataset, are then clustered using density-based OPTICS (Ankerst et al. 1999) algorithm. Finally, the sequence clusters are related to existing structural data by assessing the cluster membership of antibody sequence segments with known structure. The clusters without associated structural data are potential representatives of novel structural features.

We first tested our algorithm on a set of 10,000 artificial CDR sequences, encoded with sequence patterns mimicking those observed in real canonical classes. We clustered the artificial CDRs using the procedure described above and found that the method correctly discovered the encoded patterns. Nevertheless, a significant portion of the compressed sequences was incorrectly classified outside of the discovered sequence clusters. This tendency does not impede the correct identification of the underlying sequence patterns, as only a small number of representative sequences would be selected from each cluster for experimental characterisation.

We tested our methodology on a large antibody NGS dataset (introduced in Chapter 3) containing ~5,000,000 sequences of each chain type (κ, λ, Heavy). We extracted the CDR sequences from each chain and clustered them using the above procedure. For validation purposes, we included in the dataset human CDR sequences with known canonical class, annotated using the length-independent method described in Chapter 2. We found that the method can correctly identify most of the previously observed canonical classes and that the majority of the largest sequence clusters contain some structurally-characterised data, pointing to the conclusion that most canonical classes have already been observed. Nevertheless, we also managed to identify some relatively large sequence clusters without associated structural data, potentially representative of novel canonical classes.

The methodology can be easily adapted to process any large NGS dataset and identify sequence patterns in other protein families.

4.2 Methods

Our methodology involved projecting a large NGS dataset of CDR sequences onto a low dimensional space. In preliminary trials we tested the performance of several dimensionality reduction methods including Principal Component Analysis (PCA) (Pearson 1901), Kernel Principal Component Analysis (Kernel PCA) (Schölkopf et al. 1999), t-distributed Stochastic Neighbor Embedding (t-SNE) (Maaten et al. 2008), multidimensional scaling (Bronstein et al. 2006) and autoencoder neural networks (Hinton et al. 1994). We found that autoencoders performed best at the task of projecting the CDR sequence data.

We believe the reasons for the superior performance of the autoencoder method are as follows. First, as our objective involved clustering the projection to discover groups of structurally related CDR sequences, we anticipated that the projection should be sparse, with several dense areas in the resulting space. Out of all the methods tried, only autoencoders possessed the ability to encourage the sparsity of the projection, by using the ReLu activation function (Hahnloser et al. 2000) (see Section 4.2.2) and dropout regularisation technique (Srivastava et al. 2014) (see Section 4.2.2). Second, we found that the projection function was obfuscated in the other dimensionality-reduction methods. In contrast, in the autoencoder, by inspecting the connections in the weight matrix we could directly see which sequence patterns each neuron was detecting and qualitatively assess whether it was converging onto the expected solution, (see Section 4.2.2), Finally, neural network models have previously been used successfully at similar tasks, for example at predicting the secondary structure of proteins (Qian et al. 1988).

4.2.1 One-hot encoding of sequences

Our algorithm takes as input a set of amino acid sequences. These sequences are transformed into a matrix of numbers using one-hot encoding (Harris et al. 2013). One-hot encoding is a method of representing categorical variables as numerical vectors. If a categorical feature has N possible states, it can be encoded as a N-long vector which takes on a value of one at a position corresponding to the value of the feature and 0 at all other positions.

When applied to amino acid sequences, each amino acid can be represented as a vector of 20 positions, where each position corresponds to one amino acid type (e.g. Lysine or Glycine). In this scheme, each sequence gets transformed into a vector of length $20 \times L$, where L is the length of the amino acid sequence. The positions of the vector corresponding to the observed amino acids are encoded as 1, while the others are encoded as 0. Figure 4.1 shows a visualisation of the encoding for GFTFSTYVSY amino acid sequence.

A set of amino-acid sequences encoded using the above procedures are then used as input to our projection neural network.

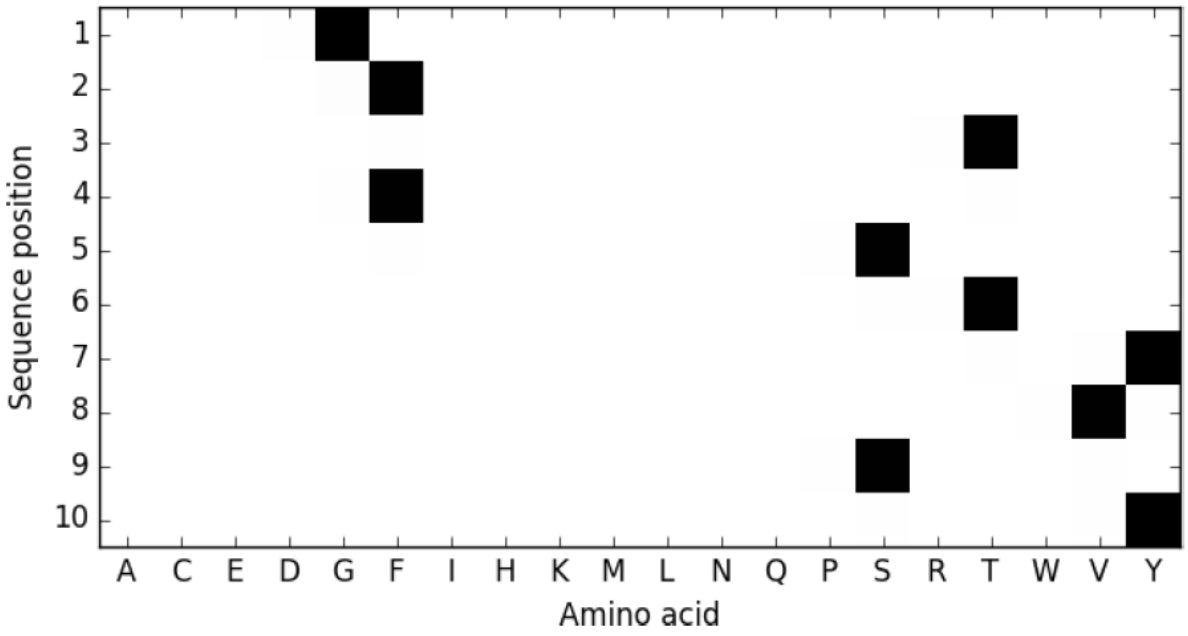

Figure 4.1 A visualisation of one-hot encoding of GFTFSTYVSY sequence. The sequence is represented in the figure as a matrix of 0s and 1s, where 1s are represented as black squares and 0s are represented as white squares. Each row corresponds to one of the positions in the sequence, and each column corresponds to an amino acid. This matrix would be flattened into a vector of length $20 \times 10 = 200$ positions where 20 is the number of proteinogenic amino acids and 10 is the number of positions in the input sequence. Such vector would then be used as an input to our algorithm.

It is important to mention here that other methods of representing amino acids exists, such as transforming sequences into vectors of physical and chemical features of constituent amino acids (Kawashima et al. 1999). We used one-hot coding to represent amino acid sequences because it is a sparse representation method which should in turn promote sparsity in the projection layer. In addition, it made the connections between the input layer and the projection layer easier to interpret, as they represented the amino acid patterns the network was recognizing. This, in turn, made it possible to inspect the

weight matrix during training, to qualitatively assess whether the network was converging to the right solution.

4.2.2 Autoencoder projection

To compress an amino acid sequence dataset, transformed into a matrix of one-hot encoded vectors, we used autoencoder neural-networks (Hinton et al. 1994). An autoencoder is a neural network machine learning algorithm, trained to reconstruct its input from a low-dimensional representation (Hinton et al. 1994).

The algorithm takes as input a matrix **X** containing the one-hot encoded sequences, with one sequence-vector per row. The input is then projected onto a hidden "compression" layer using the equation:

$$\boldsymbol{E} = \sigma(\boldsymbol{W}.\boldsymbol{X} + \boldsymbol{b}_0)$$

Where **E** is the matrix containing compressed vectors, σ is the activation function (see below), **W** is the weight matrix, $\mathbf{b}_0$ is the offset and **X** is the input matrix.

The compressed features are then reconstructed using the equation:

$$\boldsymbol{R} = \sigma(\boldsymbol{W}^T.\boldsymbol{E} + \boldsymbol{b}_1)$$

Where **R** is the reconstructed output layer, σ is the activation function (see below), **W** is the weight matrix, $\mathbf{b}_1$ is the offset and **E** is the compressed feature matrix.

For the activation function σ we used the softplus function, which is a smooth approximation to the commonly-used rectified linear unit (Hahnloser et al. 2000) activation. The softplus function is given by the following equation:

$$\sigma(x) = \ln(1 + e^{-x})$$

Where x is the input. For large positive values of x this function quickly approaches a limit of $\sigma(x) \approx x$ while for large negative values this function approaches a limit of $\sigma(x) \approx 0$. This encourages sparsity in the encoding layer, allowing different sequences to utilize only a subset of the encoding layer's dimensionality.

The weight matrix $\mathbf{W}$ and the offsets $\mathbf{b}_0$ and $\mathbf{b}_1$ are learned through gradient descent (see below) to optimize the loss L. The loss function is calculated by measuring the reconstruction error between the output layer $\mathbf{R}$ and the input layer $\mathbf{X}$.

To calculate the reconstruction error, we created a scoring function based on a symmetric version of the Blosum62 matrix (Henikoff et al. 1992). This way we can consider the chemical similarities between different amino acids. The entries of the symmetric Blosum62 matrix are calculated using the following formula:

$$\boldsymbol{B}_{i,j} = \boldsymbol{B62}_{i,j} - \frac{\boldsymbol{B62}_{i,i} + \boldsymbol{B62}_{j,j}}{2}$$

Where $\mathbf{B62}$ is the original Blosum62 matrix and $\mathbf{B}$ is the symmetric version. $\mathbf{B}$ has the properties of a distance matrix – it is symmetric and the diagonal entries are all 0. $\mathbf{B}$ is used to calculate the reconstruction error using the following formula:

$$L = \sum_{k} \sum_{i,j} \mathbf{X}_{\boldsymbol{k}}^{\boldsymbol{r}\boldsymbol{T}} . \boldsymbol{B} . \mathbf{R}_{\boldsymbol{k}}^{\boldsymbol{r}}$$

Where $\mathbf{X}_{\boldsymbol{k}}^{\boldsymbol{r}\boldsymbol{T}}$ is the transposed kth sequence-vector, reshaped into a matrix of size 20 x sequence length, $\mathbf{R}_{\boldsymbol{k}}^{\boldsymbol{r}}$ is the kth reconstructed sequence-vector, reshaped into a matrix of size 20 x sequence length. The autoencoder architecture is summarized in Figure 4.2.

The loss function is minimized using the Adaptive Moment Estimation (Adam) method (Kingma et al. 2014). The method updates the parameters of the autoencoder at each step using exponentially decaying moving averages of the mean gradients and variances.

Performing the updates using historical gradients speeds up the convergence of the algorithm (Kingma et al. 2014).

The training of the autoencoder is regularized through the commonly-used dropout technique (Srivastava et al. 2014). Dropout involves randomly occluding a fraction of the encoding layer neurons during training, so that each subset of the encoding layer is trained to reconstruct the input. This method encourages sparsity of the weight matrix, a highly desirable property (Srivastava et al. 2014).

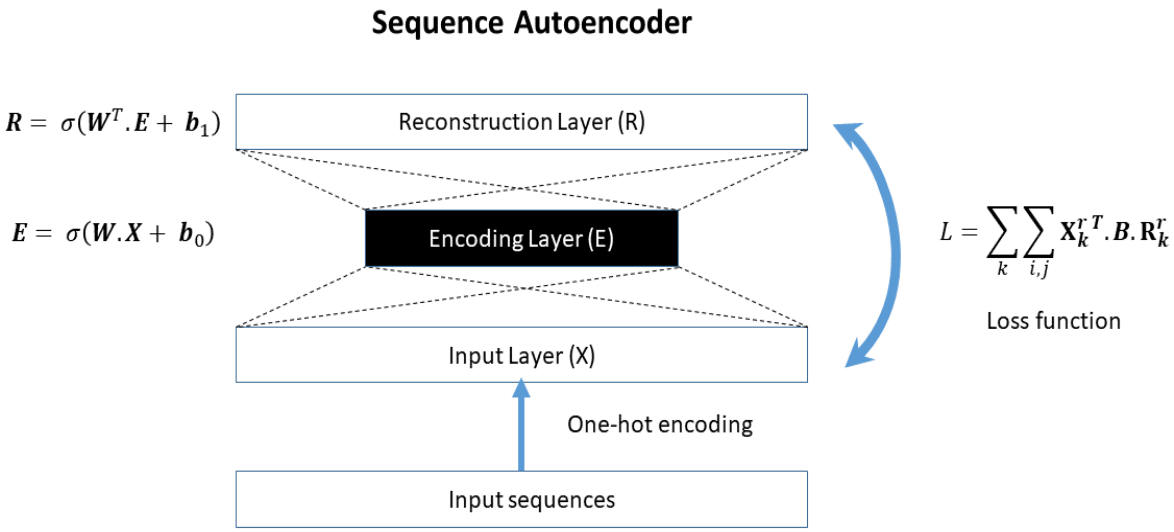

Figure 4.2 Sequence Autoencoder. A schematic of the autoencoder architecture used throughout the chapter. The input sequences are first one-hot encoded into a sparse matrix X. The matrix X is then projected onto the encoding layer E using the softplus function, compressing the input sequences. The sequences are reconstructed onto the layer R. The reconstruction error is calculated using the equation shown on the right. The meaning of the symbols shown in the figure is explained in methods Section 4.2.2.

After training of the autoencoder is completed, the compressed features are clustered using the OPTICS (Ordering Points To Identify the Clustering Structure) algorithm.

4.2.3 OPTICS clustering

The OPTICS method creates an ordering of points such that spatial neighbours are next to each other (Ankerst et al. 1999). The algorithm then calculates a "reachability distance" which specifies the local density of each point in relation to its neighbours. The reachability distances are then graphed against the ordering of points (see Figure 4.3). Dense areas of the feature space, corresponding to clusters, are represented as troughs in the OPTICS graph. The points lying outside of the dense areas are classified as "outliers" which do not belong to any clusters. As one would typically only select a small number of sequences from each cluster for experimental validation, misclassification of sequences is much more detrimental to the performance of the algorithm than incorrect placement of sequences outside of the clusters. It is therefore beneficial to use an algorithm that allows for existence of such unclassified observations. The OPTICS method is robust to non-globular shaped clusters and to clusters having variable local density (Ankerst et al. 1999). These properties are highly desirable in our work, as we found that the compressed features do not form globular clusters in the high-dimensional space and can vary in density depending on the complexity of the underlying sequence pattern.

The clustering procedure has been tested on compressed feature vectors built for CDR sequences from an antibody NGS dataset.

4.2.4 Next-Generation Sequencing dataset of CDR sequences

We tested our structural feature detection algorithm on the large NGS dataset described previously in Chapter 3. The sequences have been derived from ~500 individuals, mostly of Asian origin.

The sequences were numbered according to the Chothia scheme (Al-Lazikani et al. 1997) using the ANARCI software (Dunbar et al. 2015).

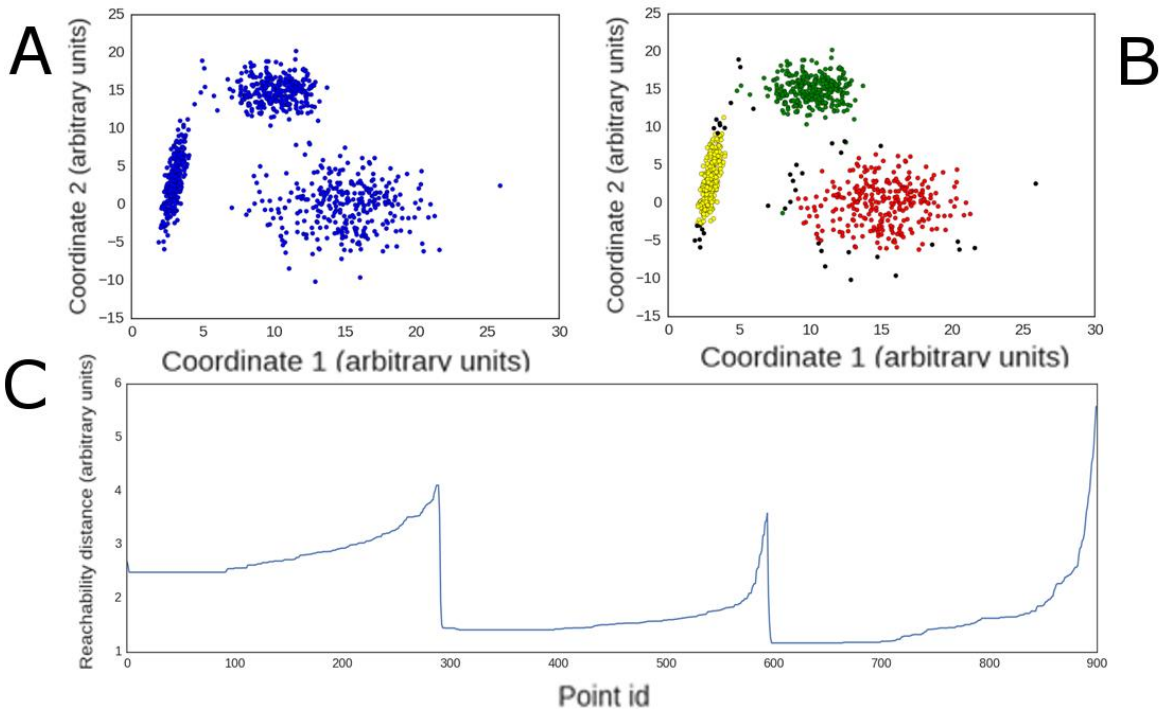

Figure 4.3 OPTICS algorithm. This figure shows the workings of the OPTICS algorithm (Ankerst et al. 1999). A) Scatter plot of 900 points with three clusters. The three clusters were sampled from 2D normal distribution with following means and covariances: 1. Mean: $\begin{pmatrix}15\\0\end{pmatrix}$ Covariance: $\begin{pmatrix}10 & 0\\0 & 10\end{pmatrix}$ 2. Mean: $\begin{pmatrix}3\\3\end{pmatrix}$ Covariance: $\begin{pmatrix}1 & 10\\0 & 10\end{pmatrix}$ 3. Mean: $\begin{pmatrix}10\\15\end{pmatrix}$ Covariance: $\begin{pmatrix}3 & 0\\0 & 3\end{pmatrix}$. B) Clusters discovered using the OPTICS algorithm. The three discovered clusters are shown in red, green and yellow. Points lying in low-density areas not classified into clusters are shown in black. C) The OPTICS plot used to find the clusters shown in panel B. The troughs in the plot correspond to the high-density areas of the scatter plot, indicating a presence of a cluster. The figure shows that the OPTICS algorithm can discover clusters of varying shapes and density.

4.2.5 CDR dataset

We extracted CDR sequences from the NGS dataset, separating them according to CDR type, chain type and CDR length (e.g. CDR-L1, κ, 11). The sequences were extracted according to CDR definitions shown in Table 4.1. These are the same definitions we used

in Chapter 2, except for CDR-H2 where we extended the definition by one residue. We modified the CDR-H2 definition because we noticed that the identity of the residue at Chothia position 49 correlates with the canonical class of the CDR-H2 loops. To each sequence set we also added sequences of CDRs from known structures (see Chapter 2) that share the same type and length.

CDR type	CDR definition
CDR-L1	24 – 34
CDR-L2	50 – 56
CDR-L3	89 – 97
CDR-H1	26 – 32
CDR-H2	49 – 56
CDR-H3	95 – 102

Table 4.1 CDR definitions. The CDR definitions presented in this table were used to extract the CDR sequences from our NGS dataset. The definitions are shown in Chothia numbering. We used the same CDR definitions in Chapter 2, except for CDR-H2 where we extended the definition by one residue.

As the NGS dataset only contains sequences of human IgMs, we only included sequences of CDR structures for which the species of origin was annotated as human (see Chapter 2). The total number of CDR sequences is given in Table 4.2. We have omitted from our analysis light chain CDR-2, because of the very low structural diversity observed for this loop type (Martin et al. 1996; Al-Lazikani et al. 1997; North et al. 2011; Nowak et al. 2016) and CDR-H3, as canonical classes have not been observed for this loop type (North et al. 2011).

	Unique	Redundant	Structures
CDR-L1	442,247	4,134,630	130
CDR-L3	585,148	4,134,630	148
CDR-K1	599,910	5,236,835	240
CDR-K3	514,416	5,236,835	288
CDR-H1	116,141	5,645,307	288
CDR-H2	164,106	5,645,307	197

Table 4.2 The number of CDR sequences. The table shows the number of sequences for each CDR type, where we made an explicit distinction between the two light chain types (where CDR-KX stands for a κ chain CDR with index X and CDR-LX stands for a λ chain with index X). The Redundant column shows the total number of CDRs extracted from our set and the Unique column shows the number of unique (at least one amino-acid difference) CDR sequences. The Structures column shows the number of unique human sequences with known structure, selected from the structural set analysed in Chapter 2. We have omitted from our analysis the light chain CDR-2 and CDR-H3.

4.3 Results & discussion

4.3.1 Overview of the procedure

We created an unsupervised machine learning method for detecting novel structural features in Next-Generation Sequencing (NGS) datasets. Throughout this chapter, we focus on analysing antibody canonical Complementarity Determining Regions (CDRs), but the algorithm could be easily adapted to process any dataset of sequence segments.

Our algorithm is based on the observation that interacting amino acid positions are correlated in a multiple sequence alignment of structurally related sequences (Morcos et al. 2011). By clustering sequences into groups with statistically correlated positions, we should therefore find groups of structurally linked observations.

The algorithm takes as input a dataset of aligned sequence segments. First, those sequence segments are encoded as sparse vectors using one-hot coding (see Section 4.2.1). Next, these vectors are used to train an autoencoder neural-network. The purpose of the autoencoder is to compress and then reconstruct the sequences from a low-dimensional sparse representation. Finally, the matrix containing the compressed sequence representations is clustered using a density-based algorithm OPTICS (Ankerst et al. 1999), which detects high-density areas in the data. The detected clusters should contain distinct sequence patterns, associated in the case of our CDR test set, with the three-dimensional shape of the corresponding structure (see Figure 4.4).

By mixing into the input dataset sequences with known structures, we can test the ability of our algorithm to associate clusters of compressed features with closely-related structures. The clusters without associated structural data could represent previously-unseen conformations.

4.3.2 Autoencoder architecture

To cluster the CDR sequence data, we designed an autoencoder neural-network which learned to reconstruct data from a low-dimensional compressed representation. In this section, we describe the details of the network architecture.

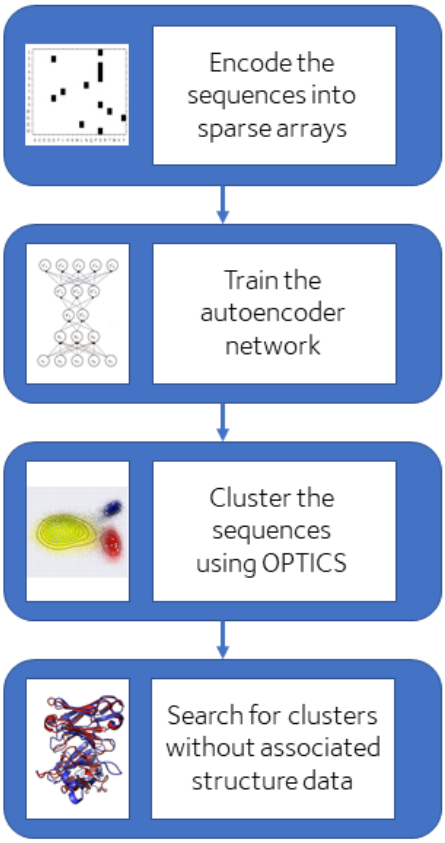

Figure 4.4 The workflow of the novel structural feature detector. The algorithm takes as input a list of sequences to be grouped into structurally-related clusters. The sequences are first encoded as a sparse array, using one-hot coding. Next, the array is used to train an autoencoder neural-network to reconstruct the input from a low-dimensional compressed representation. The compressed features are then clustered using the OPTICS (Ankerst et al. 1999) algorithm. Finally, we observe which clusters the structurally characterised sequences fall into. Clusters without associated structure data could represent novel structural features.

To assess the similarity of the reconstructed sequence to the input, we used an amino acid distance matrix. The distance matrix was built from the BLOSUM62 substitution matrix (Henikoff et al. 1992) (see Section 4.2.2). BLOSUM62 was chosen as it clusters chemically and physically related amino acids. The symmetric version of this matrix is visualized in Figure 4.5 using multidimensional scaling (Bronstein et al. 2006). The figure shows that distances between closely-related amino acids are short, while amino acids

with unique properties (Glycine, Cysteine, Proline and Histidine) fall on the fringe of the projection, far away from other residues.

We constructed our cost function using the BLOSUM62 matrix which was derived from general protein sequences contained in the BLOCKS database (Henikoff et al. 1996). We showed that the distance matrix created from this substitution table is able to capture the similarities between chemically and physically related amino acids. Nevertheless, there are other substitution models that could be used for the purpose of measuring the accuracy of sequence reconstruction. One example is the antibody-specific substitution model developed by Mirsky and co-workers (Mirsky et al. 2015). Here, we decided to use the general protein substitution model because it should make it easier in the future to extend our methodology to sequences from other protein families.

After constructing the distance matrix, we optimized the number of parameters within our network. As we did not anticipate high-order relationships to be found in the short CDR sequences, we restricted our autoencoder networks to contain a single hidden layer. We found that adding more layers did not improve the reconstruction accuracy of the autoencoder, but increased training time and made the model harder to interpret. To choose the size of the hidden layer, we measured the relationship between the reconstruction cost and the number of hidden units for hidden layer sizes between 5 and 50. The maximum size of 50 was chosen so that the total computation time did not exceed a week. We tested the networks' performance on a set of unique CDR-L1 sequences of λ type of length 13. We chose to use this CDR sequence set for testing because the CDRs of this type and length form two distinct canonical classes (see Chapter 2). The networks' performance was evaluated using four-fold validation, training the networks on three quarters of sequences and measuring the reconstruction

cost on the last quarter, for all possible training-testing set combinations. This procedure was repeated five times for each hidden layer size (see Figure 4.6). We observed constant improvement to the reconstruction cost as network size increased. In the following experiments, we set the hidden layer size to 50 neurons, which was the largest network architecture that would result in a reasonable computational time.

Next, we analysed the relationship between training time and reconstruction cost. We trained an autoencoder with 50 hidden neurons for 2,000 steps (see Figure 4.7) using the CDR-L1 test data. Initially, the training cost declines rapidly, until it reaches a level of about ~4.0 units, where the improvement becomes more incremental. We observed that the kink in the graph is created by the network reaching a "trivial" optimum, where the input sequences get reconstructed to general amino acid distribution, observed in the test CDR data. We found that training times above 1,000 steps result in very slow improvements to the reconstruction cost. We therefore set the number of training steps to 1,000.

In the next section, we describe how we used the autoencoder network to reconstruct an artificially designed test set, with pre-encoded sequence patterns.

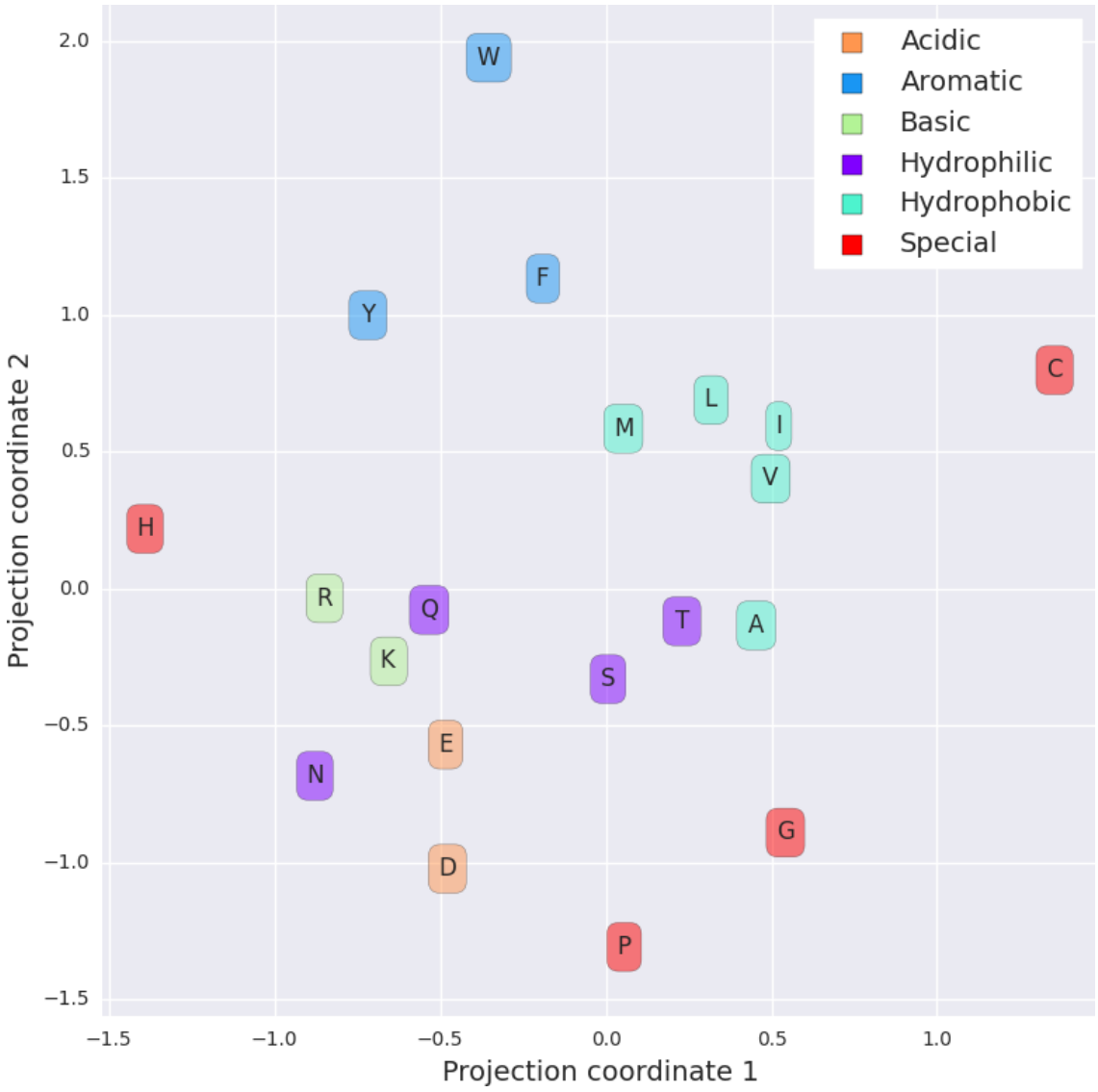

Figure 4.5 The projection of the amino-acid distance matrix. The figure shows the multidimensional scaling (Bronstein et al. 2006) projection of the distance matrix constructed from the BLOSUM62 substitution matrix. The amino acids are coloured based on their chemical properties. Acidic amino acids are coloured in orange, Aromatic amino acids are in blue, Basic amino acids are in light green, Hydrophilic amino acids are in violet, Hydrophobic amino acids are in cyan and Special amino acids are in red. The amino acids classified as "Special" have unique chemical and/or physical properties, which often prevents them from being substituted. The figure shows that physically and chemically related amino acids cluster together and that the special amino acids fall on the fringe of the projection, away from other residues.

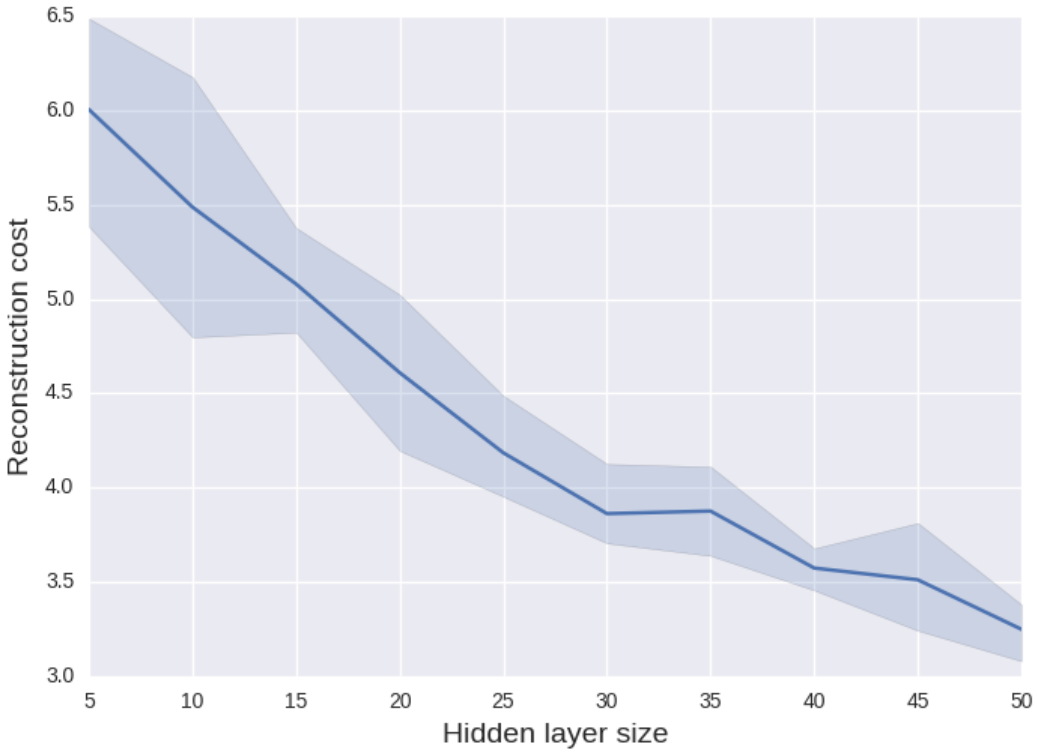

Figure 4.6 Relationship between the size of the hidden layer and the reconstruction cost. We tested the performance of our autoencoder neural-network at reconstructing a set of CDR-L1 sequences of type λ and length 13 for hidden layer sizes between 5 and 50 neurons. We tested the performance using four-fold cross-validation, training the network on three quarters of data and testing it on the fourth. The networks were trained for 500 steps. We calculated five repeats for each network size. The shaded region represents a 95% confidence interval, calculated using bootstrapping.

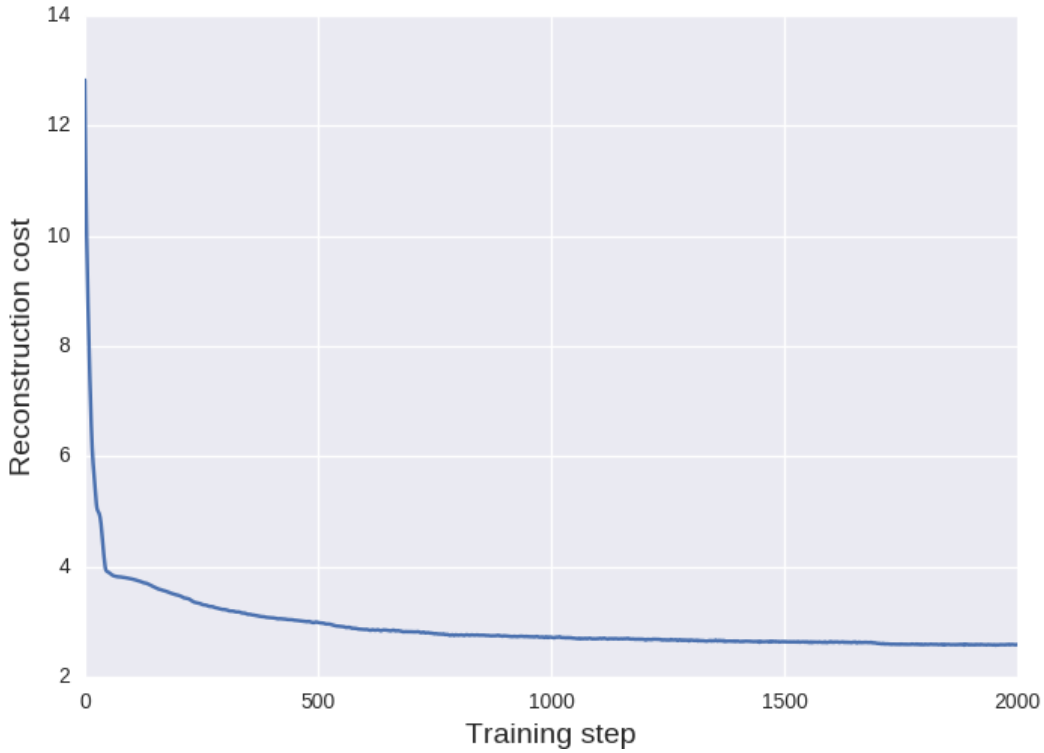

Figure 4.7 The relationship between reconstruction cost and training time. The graph was calculated for an autoencoder with 50 hidden neurons, tasked with reconstructing a set of CDR-L1 sequences of type λ and length 13. Initially, the reconstruction cost declined rapidly, until it reached a level of about 4.0, where the improvement became more gradual. We found that the increase in performance became very slow after 1,000 steps.

4.3.3 Artificial CDR dataset validation

Having optimized the autoencoder architecture, we tested the algorithm's performance at discovering amino-acid correlations in an artificially-designed CDR dataset.

The set contained 50,000 artificial amino-acid sequences of length seven. We encoded two patterns into the set. The first pattern consisted of a Hydrophilic residue at position 2 (Serine or Threonine), a Hydrophobic residue at position 4 (Leucine or Isoleucine) and a negatively charged amino acid at position 6 (Glutamic acid or Aspartic acid). The second pattern contained a Glycine at position 1 and a Proline at position 4. These patterns are

rough representations of real correlations observed in CDR canonical classes (see Chapter 2). The positions not included in the definition of a pattern contained amino acids sampled from a categorical distribution representing the set of 20 amino acids, according to probabilities sampled from a Dirichlet distribution (Kotz et al. 2000) with all concentration parameters set to 0.5 (non-informative Jeffrey's prior (Jeffreys 1946)). The first 24,500 sequences in the artificial CDR dataset were formed using the first pattern, the next 24,500 were coded using the second pattern and the final 1,000 were sampled from a uniform distribution (see Figure 4.8). The sequences sampled from the uniform distribution represent noise and constitute 2% of the whole artificial CDR dataset.

We encoded the sequences and trained the autoencoder for 1,000 training steps. The compressed sequences were then clustered using the OPTICS algorithm. The results are shown in Figure 4.9. We found that the clusters discovered using the OPTICS algorithm accurately reproduce patterns encoded into our artificial CDR dataset. The confusion matrix created for the clustering is shown in Table 4.3. To quantify the quality of our clustering we calculated the homogeneity metric, which measures to what extent each cluster contains sequences from only a single pattern. The metric is bound between 0.0 (random correspondence) and 1.0 (perfect correspondence). We obtained a homogeneity of 0.63, showing a good correspondence between the original patterns and the discovered clusters.

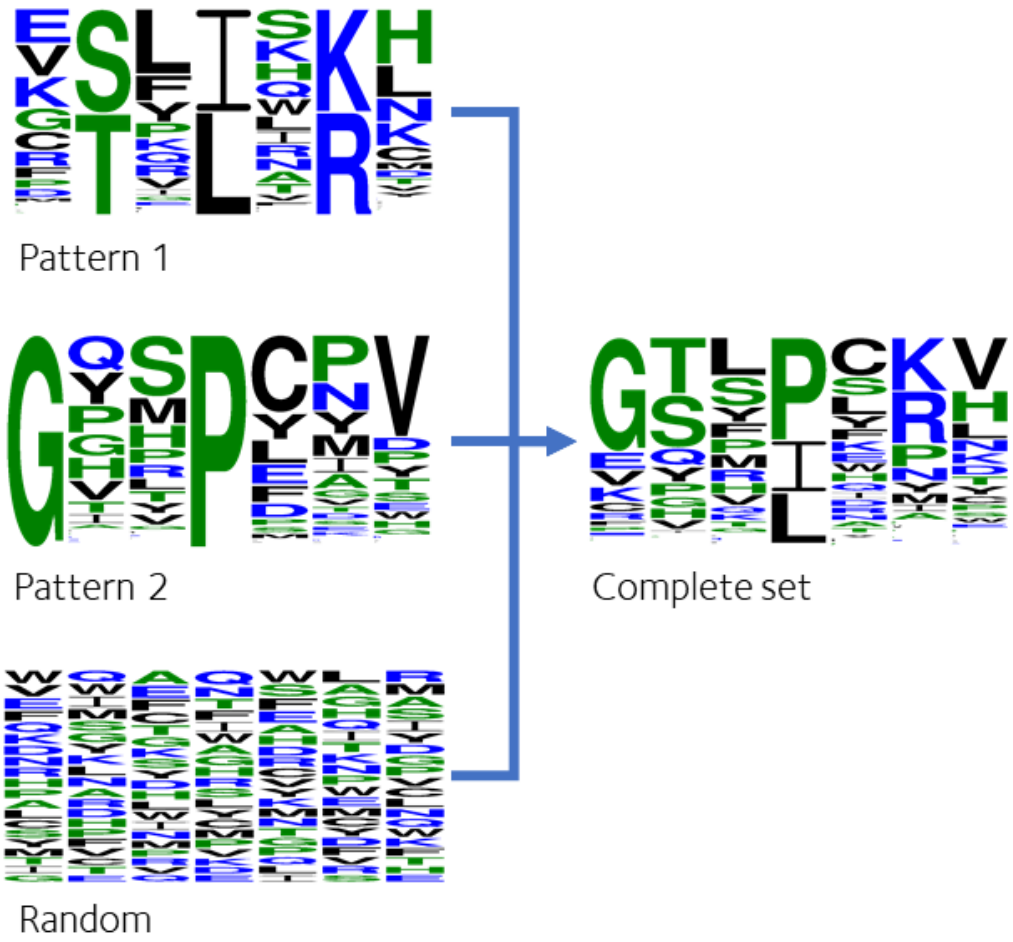

Figure 4.8 The artificial CDR dataset. The figure shows sequence logos created for amino-acid patterns encoded within our artificial CDR dataset of 50,000 simulated sequences of length 7. The dataset contained two patterns, each consisting of 24,500 sequences. Pattern 1 consisted of Serine or Threonine at position 2, Leucine or Isoleucine at position 4 and Arginine or Lysine at position 6. Pattern 2 consisted of a Glycine at position 1 and a Proline at position 3. The final 1,000 sequences in the set were sampled from a uniform distribution and represent noise.

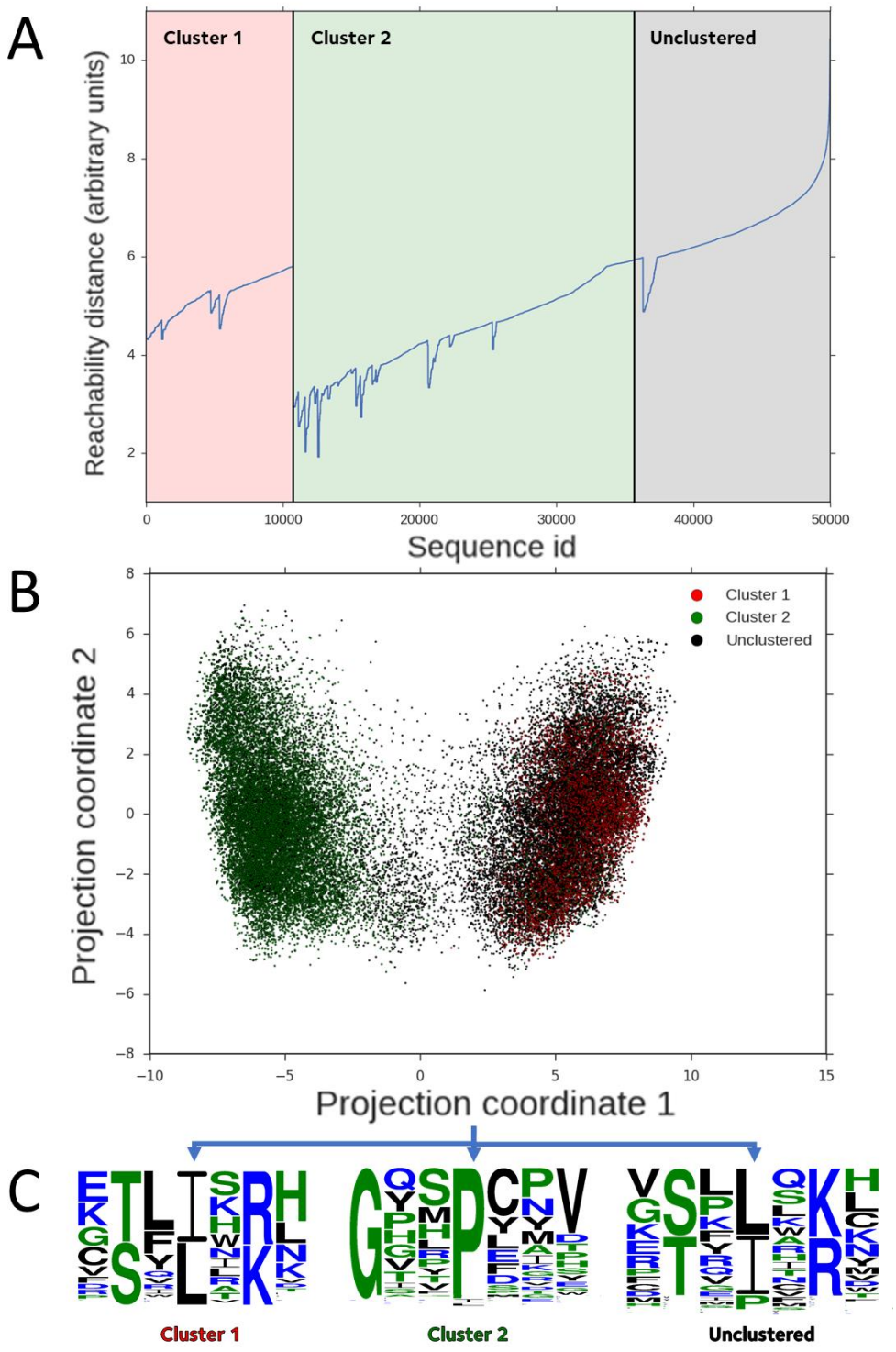

Figure 4.9 The clustering of the artificial CDR dataset. A) The OPTICS plot calculated for the compressed sequence features, calculated using the autoencoder network. The shaded regions correspond to the discovered clusters. Cluster 1 is shown in red, Cluster 2 is shown in green and Unclustered region is shown in grey. B) Principal Component Analysis (PCA) of the 50-

dimentional compressed feature matrix. Each point in the plot corresponds to one of the sequences in our test dataset. The points have been coloured according to their cluster membership. Points belonging to Cluster 1 are shown in red, points belonging to Cluster 2 are shown in green and unclustered points are shown in black. C) Sequence logos created for sequences from each discovered cluster. Our clustering method reproduced the patterns encoded into the artificial CDR dataset (see Figure 4.8).

We observed that many of the sequences from Pattern 1 were incorrectly placed outside of the clusters. This is likely related to the complexity of the sequence pattern, as more complex patterns produce more "diluted" compressed features. This effect can be seen in OPTICS plot shown in Figure 4.9A where the reachability distances are higher for members of Cluster 1 (red shaded area) than for Cluster 2 (green shaded area). This will place some of the sequences in low-density areas, preventing accurate classification. Nevertheless, we found that only a small fraction (~2%) of sequences were placed in an incorrect cluster, demonstrating the power of the methodology. As one would usually select only one or two sequences from each cluster for structural characterisation, incorrect classification of sequences as lying outside of the main clusters does not prevent the correct identification of the constituent sequence patterns.

	Pattern 1	Pattern 2	Random
Cluster 1	10,460	1	3
Cluster 2	1,316	23,001	247
Unclustered	12,724	1,498	750

Table 4.3 The confusion matrix for the classification of the artificial CDR dataset. The columns in the table show the number of sequences for each pattern (24,500 for Pattern 1 and 2 and 1,000 for Random). The rows show the number of sequences classified into each cluster and the number of sequences considered to lie outside of the clusters. The correct assignments are highlighted in green, while the incorrect ones are highlighted in red.

In this section, we conducted a preliminary test on a set of artificial sequences. In the test set, we empirically encoded patterns similar to those observed in real canonical classes, found in Chapter 2. While such simple, preliminary test provides only a limited approximation to the NGS set of real CDR sequences, the results provide some indication that our autoencoder network can differentiate between groups of sequences with correlated positions. Motivated by this result, we conducted a large-scale analysis on a dataset of real CDR sequences. In the next section, we show the results of this analysis and cross-validate the clusterings with the canonical forms described in Chapter 2.

4.3.4 CDR canonical classes validation

We have shown the power of our algorithm on a dataset of simulated CDR sequences. In this section, we searched for patterns in datasets of real CDR sequences (see Section 4.2.5) and compared the clusterings with the CDR canonical class definitions calculated in Chapter 2, which we considered to be ground truth.

For the purpose of validating our method against the CDR canonical classes, we selected groups of structurally characterised CDR sequences, assembled based on CDR sequence length, chain type and CDR type, containing at least 6 unique human sequences with known crystal structure. The sequence groups fulfilling this criterion were: CDR-K1 Length 11, CDR-K1 Length 15, CDR-K1 Length 16, CDR-K1 Length 17, CDR-L1 Length 11, CDR-L1 Length 13, CDR-L1 Length 14, CDR-K3 Length 8, CDR-K3 Length 9, CDR-L3 Length 9, CDR-L3 Length 11, CDR-H1 Length 7, CDR-H2 Length 7 and CDR-H2 Length 8. Next, we mixed those groups of structurally characterised sequences with the corresponding NGS sequence groups (see methods section 4.2.5). We then trained autoencoder networks with 50 hidden neurons for 1,000 steps, separately for each combined group of sequences. The results are shown in Table 4.4. We found that if a CDR type forms only a single canonical class (e.g. CDR-K1 Length 11 or CDR-K3 Length 9) our algorithm usually grouped over 90% of sequences into a single sequence cluster (e.g. Cluster-3 for CDR-K1 Length 11), with the rest of sequences falling into a couple of smaller clusters.

CDR-K1 Length 11

	Cluster-1	Cluster-2	Cluster-3	Unclustered
No. of sequences	3,072	4,032	201,805	7,863
L1-10,11,12-A	1	4	83	4
Not in a large canonical class	0	0	1	2

CDR-K1 Length 15

	Cluster-1	Unclustered
No. of sequences	4,210	1,166
L1-15-A	2	9
Not in a large canonical class	0	2

CDR-K1 Length 16

	Cluster-1	Cluster-2	Unclustered
No. of sequences	35,927	35,746	10,630
L1-16-A	3	11	0
Not in a large canonical class	1	0	2

CDR-K1 Length 17

	Cluster-1	Cluster-2	Unclustered
No. of sequences	178,275	17,852	19,032
L1-17-A	10	0	0
Not in a large canonical class	0	0	0

CDR-L1 Length 11

	Cluster-1	Cluster-2	Cluster-3	Unclustered
No. of sequences	14,879	3,085	2,540	66,486
L1-11-A	3	0	0	6
L1-11-B	0	1	0	5
Not in a large canonical class	1	0	0	9

CDR-L1 Length 13

	Cluster-1	Cluster-2	Unclustered
No. of sequences	95,968	5,363	32,132
L1-13,14-A	20	0	2
L1-13-A	0	6	0
Not in a large canonical class	1	0	0

CDR-L1 Length 14

	Cluster-1	Cluster-2	Cluster-3	Unclustered
No. of sequences	151,496	21,075	4,215	27,830
L1-13,14-A	7	4	0	2
Not in a large canonical class	1	0	0	3

CDR-K3 Length 8

	Cluster-1	Unclustered
No. of sequences	50,564	5,106
L3-8-A	9	0
Not in a large canonical class	4	1

CDR-K3 Length 9

	Cluster-1	Cluster-2	Cluster-3	Cluster-4	Unclustered
No. of sequences	232,863	10,585	11,996	5,646	44,554
L3-9,10-A	113	2	4	2	15
Not in a large canonical class	7	0	0	0	4

CDR-L3 Length 9

	Cluster-1	Cluster-2	Cluster-3	Unclustered
No. of sequences	43,548	3,871	968	6,846
L3-9-A	5	0	0	1
Not in a large canonical class	5	0	0	1

CDR-L3 Length 11

	Cluster-1	Cluster-2	Cluster-3	Unclustered
No. of sequences	4,032	33,770	106,351	90,259
L3-10,11-A	0	0	18	3
Not in a large canonical class	0	2	0	8

CDR-H1 Length 7

	Cluster-1	Unclustered
No. of sequences	57,823	1,615
H1-7-A	140	3
Not in a large canonical class	24	0

CDR-H2 Length 8

	Cluster-1	Unclustered
No. of sequences	43,226	19,390
H2-7-A	21	1
Not in a large canonical class	3	4

CDR-H2 Length 9

	Cluster-1	Cluster-2	Unclustered
No. of sequences	20,161	139,617	82,414
H2-8-A	27	6	53
H2-8-B	0	37	10
Not in a large canonical class	2	8	5

Table 4.4 Clustering results for real CDR sequences. The table shows the clustering results for CDR sequence groups containing at least 6 unique, human, structurally-characterized sequences. The "No. of sequences" row shows the number of sequences classified into each discovered cluster. The following rows correspond to canonical classes containing at least 6 unique human sequences and show the number of structurally characterised sequences falling into each discovered cluster. For the three CDR types containing more than one canonical class, we have highlighted the correctly classified sequences in green and incorrectly classified sequences in red.

In general, we have found very few cases where structurally distinct loops have been classified to the same cluster. This shows that our method correctly identified the sequence patterns underlying the known canonical classes. Nevertheless, in several cases, we found that the structurally characterised CDRs from a single canonical class have been split over a number of sequence clusters. We found that this is usually caused by a single structural shape being coded for by a number of different sequence patterns, usually related to the underlying germline genes. This principle is illustrated in Figure 4.10

with the example of CDR-K1 loops of length 16. We found two sequence clusters for this CDR-type containing approximately the same proportion of the sequence data. Nevertheless, only one canonical class has been identified for this loop type (L1-16-A) and the structurally characterised sequences are split between the two clusters. Inspecting the constituent sequences, we find that the clusters are split by the identity of the 10th amino acid in the CDR sequence, which is Aspartic acid in Cluster-1 and Asparagine in Cluster-2. We find that, in most cases, the identity of this amino acid is determined by the variable germline gene coding for the CDR sequence. Most of CDRs in Cluster-1 are coded for by IGKV2-30 or IGKV2-24, containing an Aspartic acid at position 10 in the sequence while most of CDRs in Cluster-2 are coded for by IGKV2-28 which contains an Asparagine at this position. Our algorithm separated the amino acid patterns of the different germlines. This example shows that there are cases where different amino acid patterns do not correspond to distinct loop structures. The amino acid patterns in the sequence clusters detected by our algorithm should therefore be manually inspected to determine whether they could be coding for the same structure.

There were three groups of CDR sequences, containing sufficient structural data, with more than one corresponding canonical class. Those were CDR-L1 length 11, CDR-L1 length 13 and CDR-H2 length 9. For these groups, we have analysed the relationship between the structural canonical classes and the sequence clusters in detail. Comparisons between sequence logos created for the NGS CDR sequence clusters and for corresponding canonical classes are shown in Figure 4.11.

We found (see Table 4.4) that most structurally characterised CDR-L1 loops of length 11 were not assigned into CDR sequence clusters. We showed in Section 4.3.3 that as the complexity of the pattern increases, the coverage drops, which could explain the

observed results, as the split between the CDR-L1 length 11 canonical classes is related to several amino acid interactions (see Chapter 2). Despite low coverage of the structural data, the sequence patterns present in the clusters replicate those observed in canonical classes (see Figure 4.11).

Our algorithm correctly separated the two canonical classes of CDR-L1 loops of length 13 (see Table 4.4). The discovered sequence clusters contained virtually the same sequence patterns as the corresponding canonical classes (see Figure 4.11). This result shows the power of our method, when the structural differences have a strong foundation in the underlying residue relationships.

Two sequence clusters were identified for CDR-H2 loops of length 9. For the purposes of the analysis conducted in this chapter we extended the CDR-H2 definition to include residue at Chothia position 49, as we found that the identity of this residue correlates with the canonical class of the loops of this type (see Section 4.2.5). The H2-8-A canonical class is characterised by a Glycine at position 1, Proline at position 5 and Glycine at position 8. H2-8-B is characterised by Serine/Alanine at position 1, Serine at position 4 and a Glycine at position 7 (see Figure 4.11). The sequence clusters discovered by our method replicate those sequence patterns and correctly separate most of the structurally characterised loops (see Table 4.4).

In this section, the performance of our algorithm was tested at reproducing the sequence patterns observed in CDR canonical classes. We have observed that, at times, the method split the sequences based on underlying genetic data, instead of structural relationships. Nonetheless, the clusters discovered by the algorithm accurately reproduced the sequence patterns of the CDR canonical classes. In the next section, we

analyse CDR sequence data with low structural coverage, to identify potential novel canonical classes.

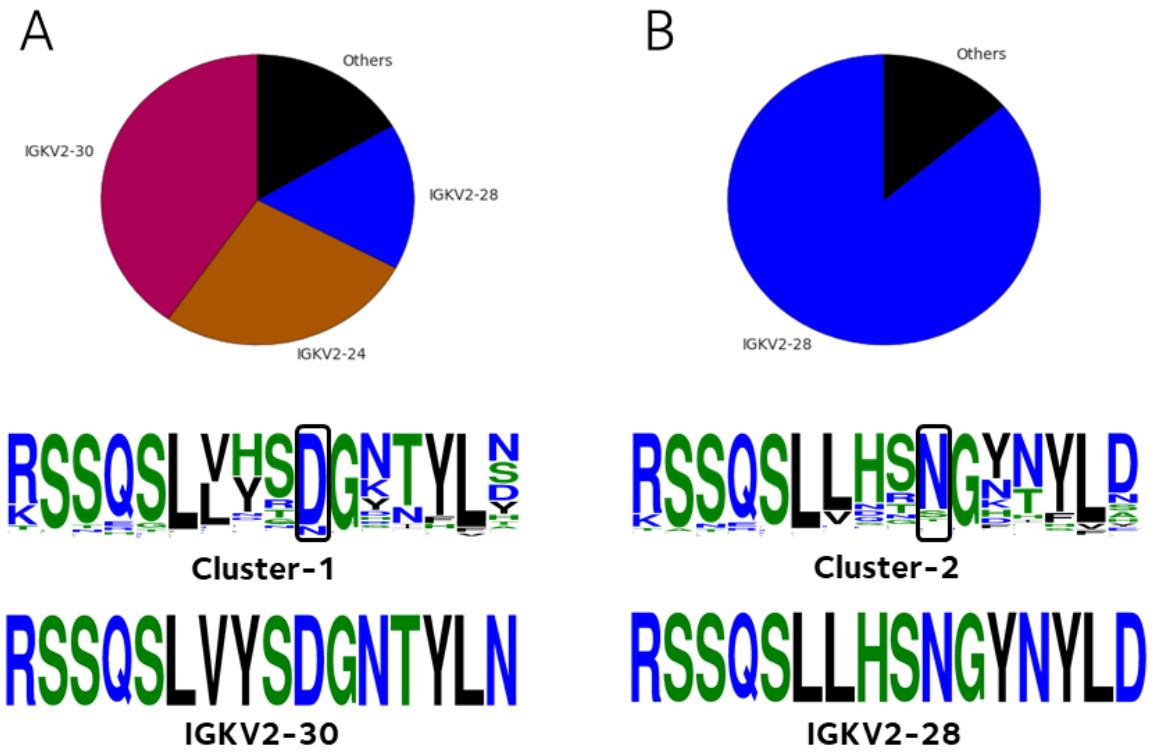

Figure 4.10 Two clusters of CDR-K1 loops of length 16. The top row contains pie charts showing the distribution of germline genes coding for loops contained within the sequence clusters of CDR-K1 loops of length 16. Panel A shows the pie chart for Cluster-1 and Panel B shows the pie chart for Cluster-2. The second row contains the sequence logos, showing the distribution of amino acids, at each position for CDR sequences contained within Cluster-1 (Panel A) and Cluster-2 (Panel B). The bottom row shows the sequence of the most common germline coding for CDRs within Cluster-1 (IGKV2-30, Panel A) and Cluster-2 (IGKV2-28, Panel B). The image shows that our algorithm is splitting the CDR-K1 sequences based on the identity of the shared amino acid at position 10 (outlined in the figure for both clusters) which is determined by the corresponding germline. The amino acid at position 10 is Aspartic acid in Cluster-1 and Asparagine in Cluster-2.

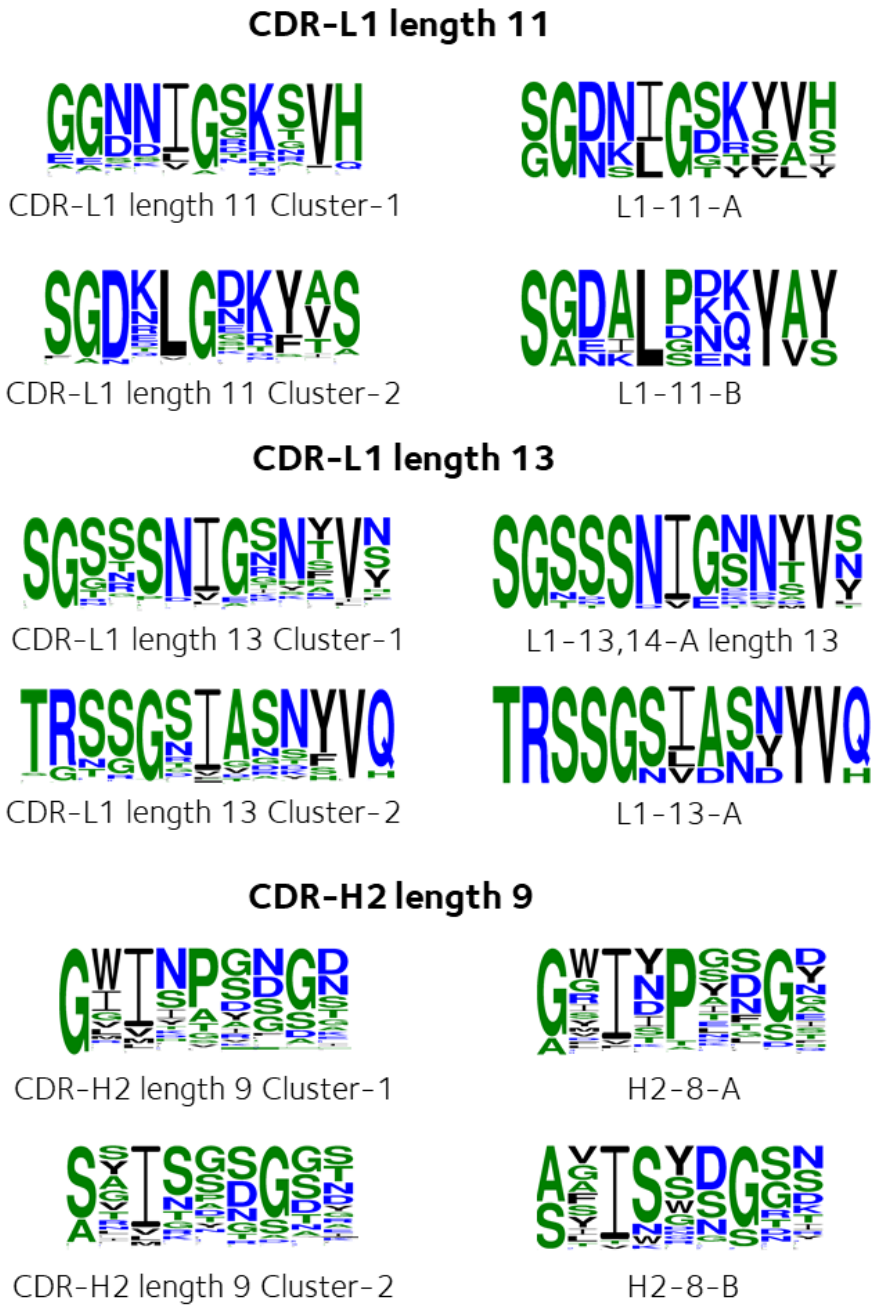

Figure 4.11 Sequence logo comparison. The figure compares sequence patterns of canonical classes and the corresponding NGS CDR sequence clusters, discovered using our algorithm. Sequence logos created for NGS CDR sequence clusters are shown on the left, while sequence logos for corresponding canonical classes are shown on the right. Only sequences classified as human were used to plot the canonical class sequence logos. The comparisons are shown for three groups of CDR sequences containing sufficient structural data for which two canonical

classes have been described in Chapter 2. These CDR groups were CDR-L1 length 11, CDR-L1 length 13 and CDR-H2 length 9.

4.3.5 Novel sequence patterns

We searched the NGS CDR sequence data for novel sequence patterns, not represented in the structural data. Such novel patterns could be related to previously unseen CDR structures and correspond to novel canonical classes. Our collaborators at UCB Pharma are currently expressing and structurally characterising a small number of Fv sequences containing these interesting patterns.

Towards discovering novel canonical classes, we have trained autoencoder networks on all CDR sequence groups containing over 1,000 unique sequences. We clustered the compressed sequence features using the OPTICS algorithm and manually inspected sequence patterns in the discovered clusters. In this section, we describe three previously uncharacterised CDR sequence clusters, which potentially represent novel canonical classes. These patterns were found for CDR-L3 length 13, CDR-L1 length 14 and CDR-H1 length 10.

4.3.5.1 CDR-L3 length 13

There were 20,842 unique NGS sequences of this type in our dataset, but only three sequences of known structure. This suggests this CDR type is under represented in the structural data. The clustering results obtained for the NGS data are shown in Figure 4.12 along with corresponding germline distributions. There are three clusters with distinct amino acid patterns. All structurally characterised sequences lay within Cluster-2. Through manual inspection of the structures, we observed that the shape was preserved by the Tryptophan at position three, Aspartic acid at position four and Serine at position six (see Figure 4.13). The sequence pattern of Cluster-3 contained both residues, so we

considered it unlikely that CDRs within this cluster would have a significantly different shape to those present in Cluster-2. Cluster-1 contains loops coded for by the IGLV9-49 germline and the sequences contained within this cluster do not seem to have the capacity to form interactions observed in the structurally characterised loops. The relative scarcity of CDRs produced by this germline can be explained by IGLV9-49 containing a stop codon at the C-terminus of the sequence. Therefore, for this sequence to be expressed, the stop codon must be removed during the VJ recombination.

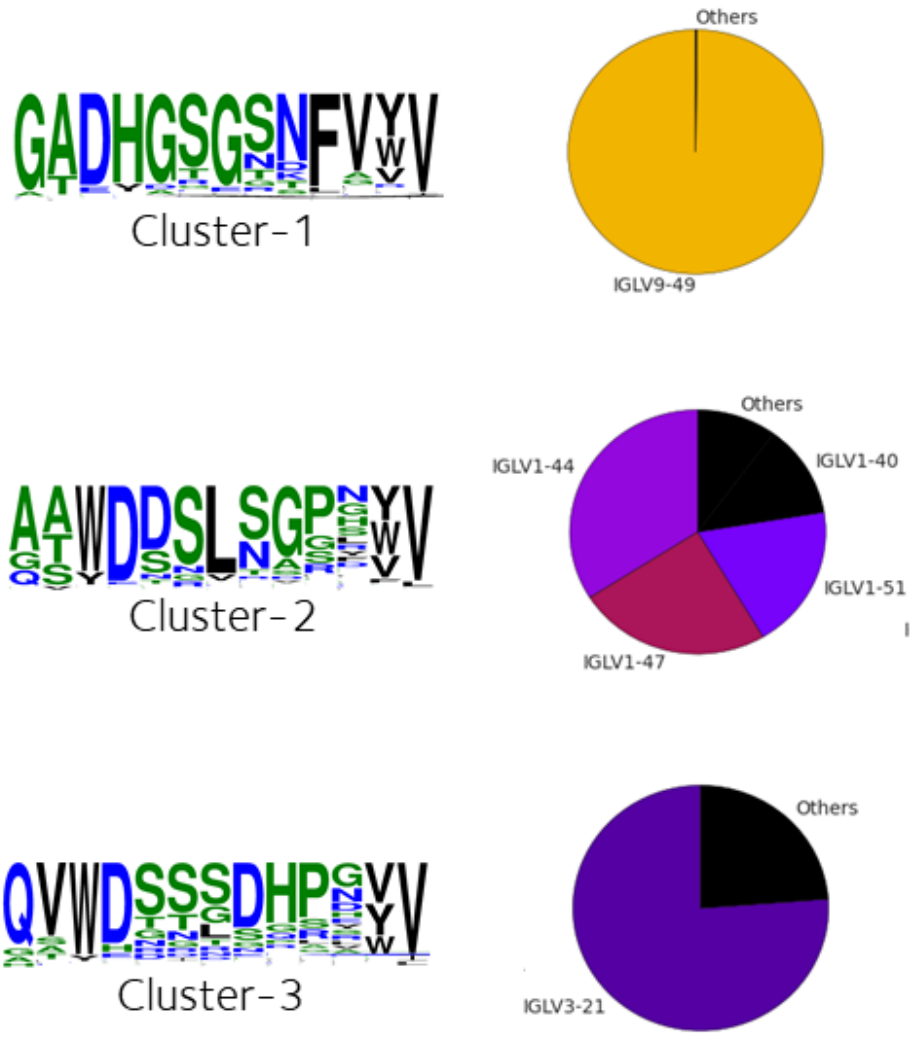

Figure 4.12 CDR-L3 length 13 clusters. Sequence logos, along with the germline distributions are shown for the three sequence clusters found for CDR-L3 loops of length 13. Three loops with known structure were classified into Cluster-2. The loops in Cluster-1 were predominantly coded for by the IGLV9-49 germline, loops in Cluster-2 were coded for by germlines from the IGLV1 subgroup and loops in Cluster-3 were coded for by IGLV3-21 germline.

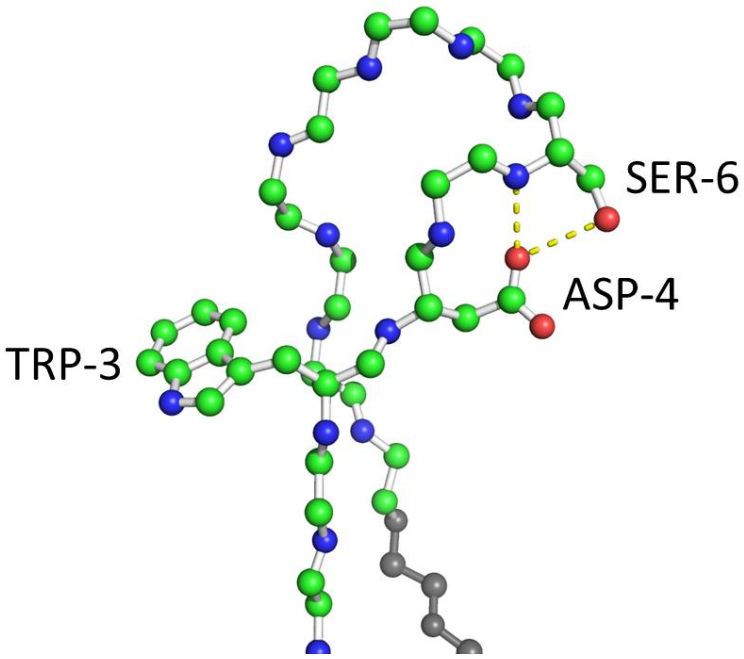

Figure 4.13 The interactions responsible for the shape of CDR-L3 loops of length 13 with known structure. The schematic shows the CDR-L3 loop of antibody with PDB id 3MLW. The sequence of the loop is ASWDDSRGGPDYV. The Carbon atoms are shown in green, the Nitrogen atoms are shown in blue and the Oxygen atoms are shown in red. The residues outside of the Chothia CDR-L3 definition are shown in grey. The structure of loop is presented in ball and stick representation showing the backbone atoms of the constituent residues and the side chains of three residues likely responsible for the structure of the loop (Tryptophan at position three, Asparagine at position four and Serine at position six). The hydrogen bonds between Asparagine at position four and Serine at position six are shown using yellow dashed lines. The sequence of this loop was classified into Cluster-2 (see Figure 4.12).

4.3.5.2 CDR-L1 length 14

There were 204,616 unique NGS sequences and 18 structurally characterised CDR-L1 loops of length 14 in our dataset (see Table 4.4). We found three clusters for loops of this type. The sequence logos, along with corresponding germline distributions, are shown in Figure 4.14. The structurally characterised loops from L1-13,14-A canonical class were classified into Cluster-1 and Cluster-2. We previously observed (see Chapter 2) that the L1-13,14-A structures are stabilized by the Aspartic acid/Asparagine at position six in the CDR sequence forming hydrogen bonds with neighbouring residues. Cluster-3 did not contain any human structural loops and had a unique sequence pattern which we considered unlikely to preserve the interactions observed in L1-13,14-A. Using the Protein BLAST (Altschul et al. 1997) webserver, we identified three structurally characterised CDRs from the Rhesus Monkey that are similar to the sequences in Cluster-3. These Rhesus Monkey CDRs have been described in the literature as having a unique binding mode to the HIV virus (Tran et al. 2014).

The Rhesus Monkey CDRs also take on a very different shape to the structurally characterised human CDRs. It would be a great validation of our work if the human sequences in Cluster-3 were shown to have a structure similar to the Rhesus Monkey CDRs (this case is currently being tested by UCB).

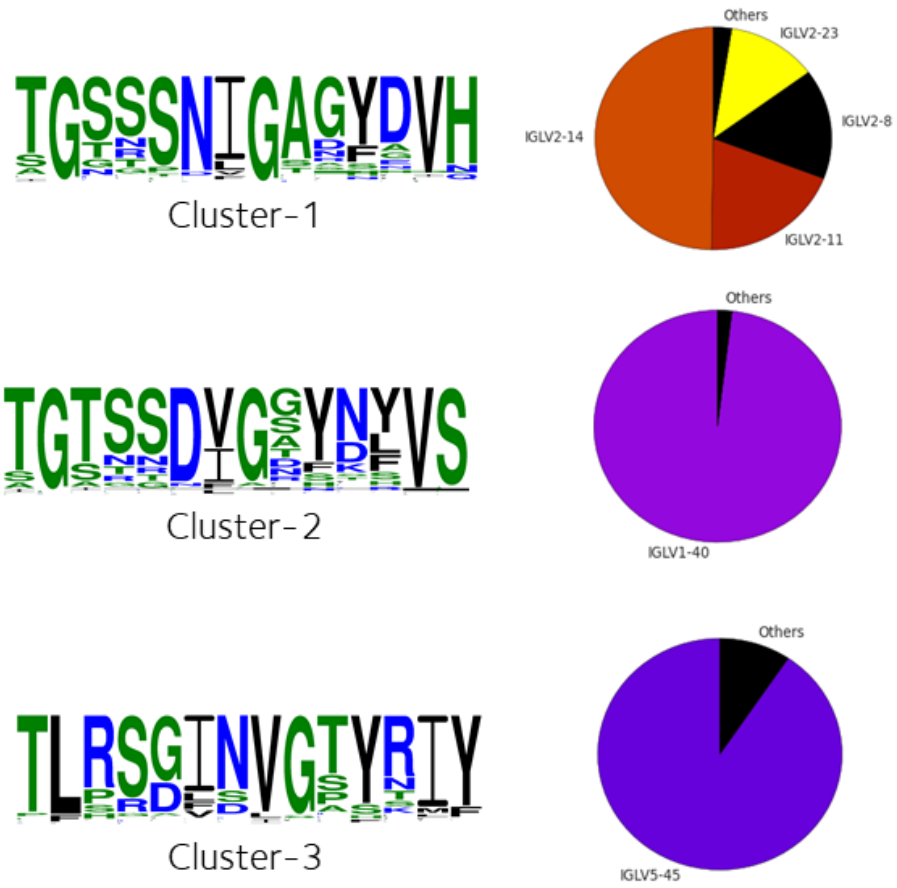

Figure 4.14 CDR-L1 length 14 clusters. Sequence logos, along with the germline distributions are shown for three CDR sequence clusters found for CDR-L1 loops of length 14. The structurally characterised human loops were classified into Clusters 1 and 2 (see Table 4.4). The loops in Cluster-1 were predominantly coded for by the germlines from the IGLV2 subgroup, loops in Cluster-2 were coded for by the IGLV1-40 germline and loops in Cluster-3 were coded for by the IGLV5-45 germline. The sequence pattern observed in Cluster-3 suggests that CDRs contained in this cluster would not be able to form the interactions responsible for the structural shape of CDRs classified into Clusters 1 and 2, suggesting the sequence pattern might create a novel canonical class.

4.3.5.3 CDR-H1 length 10

No human structurally-characterised sequences of this type were present in our dataset. The NGS data contained 1,194 CDR sequences of this type. The majority of the human CDR-H1 loops were of length 7 and 9. The structures of these loops were preserved by hydrophobic residues present at the second and fourth position in the sequence (see Chapter 2). Inspecting the sequence patterns for clusters discovered for CDR-H1 loops of length 10 we find that the loops in Cluster-1 and Cluster-2 are unlikely to contain the interactions observed in the shorter CDR-H1s (see Figure 4.15). Structural characterisation of antibodies containing CDR-H1 loops of length 10 could reveal previously unseen binding site shapes.

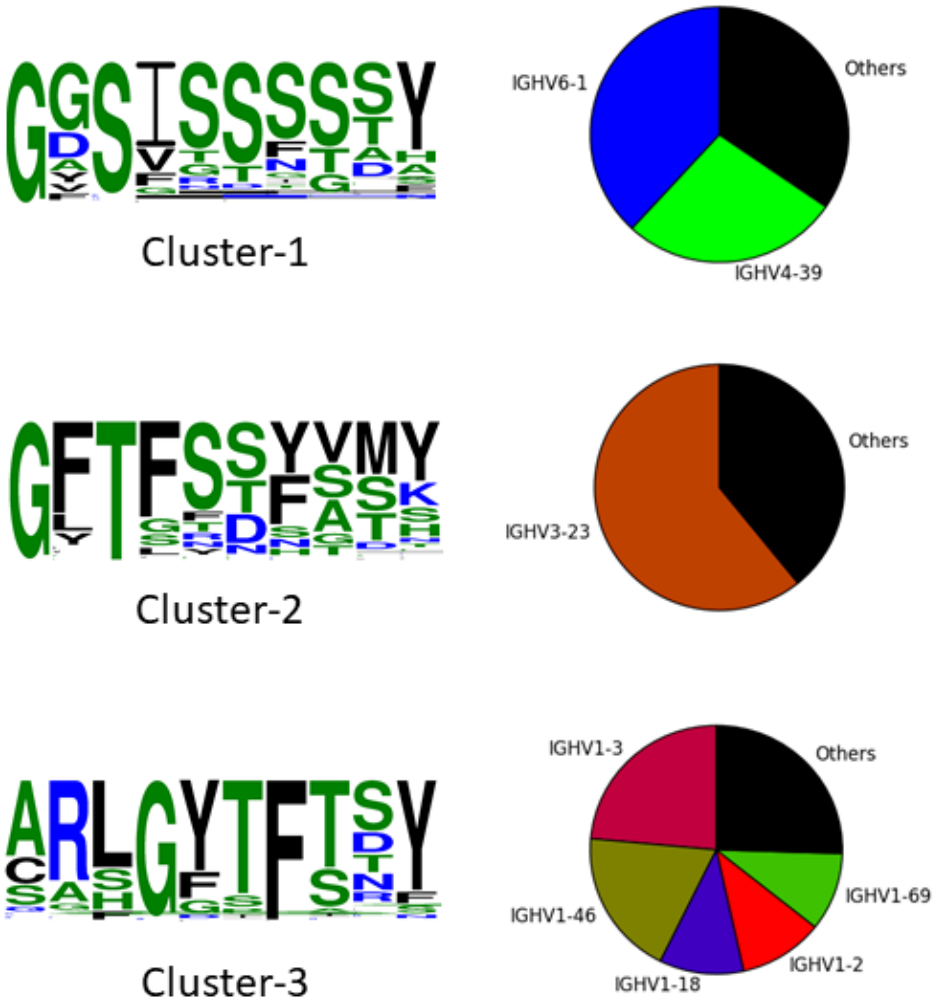

Figure 4.15 CDR-H1 length 10 clusters. Sequence logos, along with the germline distributions are shown for three CDR sequence clusters found for CDR-H1 loops of length 10. No structurally-characterised CDR sequences were found for this loop type and length. The loops in Cluster-1 were predominantly coded for by the IGHV6-1 and IGHV4-39 germlines, loops in Cluster-2 were coded for by the IGHV3-23 germline and loops in Cluster-3 were mostly coded for by germlines from the IGHV1 subgroup. The sequence patterns observed in Clusters 1 and 3 differ significantly from the patterns observed in CDR-H1 loops of known structure, suggesting they could take on previously unseen conformations.

4.4 Conclusions

In this chapter, we created a machine learning method for finding groups of related sequences within large Next-Generation Sequencing (NGS) datasets of antibody CDR sequences. The algorithm used autoencoder neural-networks to project the sequence data onto a low-dimensional representation. Sequences forming related sequence patterns should be projected onto similar representations, forming clusters of features. To discover the clusters, the algorithm used density-based clustering algorithm OPTICS. The structural significance of the discovered clusters was assessed by associating the sequence patterns with the existing structural data.

First, we tested our algorithm on an artificial set of 10,000 amino acid sequences, containing sequence patterns mimicking those observed for CDR canonical classes in Chapter 2. We trained an autoencoder on the artificial dataset and clustered the resulting projection. The algorithm correctly identified the sequence patterns imputed into the dataset, despite many sequences being placed outside of the main clusters. This propensity to classify sequences outside of the main clusters did not hinder correct identification of the underlying sequence patterns and only a small number of sequences would be selected from each cluster for further experimental study.

Next, we validated our method on groups of CDR sequences from a large NGS dataset, containing over 15,000,000 sequences of single, human IgM chains. The discovered clusters were compared to the sequence patterns observed for CDR canonical classes in Chapter 2. We found good correspondence between the canonical classes and the discovered sequence clusters. Nevertheless, in some cases the patterns discovered by our algorithm were related to the underlying biological processes responsible for antibody generation, instead of structure-stabilizing inter-residue interactions. This

result shows that one needs to consider the functionality of the analysed protein sequences to interpret the results correctly. We analysed in detail three groups of CDR sequences for which more than one canonical class was observed in Chapter 2. We found that our algorithm correctly identified the sequence patterns observed in canonical classes. This result demonstrated the power of our method at clustering NGS datasets based on structurally relevant amino acid correlations.

In the final section, we presented three examples of potential novel canonical classes. These exotic sequence clusters were coded for by rarely-used germlines and consisted of sequence patterns which were unlikely to preserve the interactions observed in structural data. Our collaborators at UCB Pharma are currently expressing and structurally characterising these novel patterns. Structural characterisation of CDRs created by these patterns could lead to discovery of novel canonical classes and enrich our understanding of antibody structural space.

5 DATA-DRIVEN ANTIBODY DESIGN

5.1 Introduction

Since the first monoclonal antibody was synthesized in 1975 the field of antibody therapeutics has grown exponentially (The sales of monoclonal antibodies reached $20.6 billion (Maggon 2007) in 2006). Nowadays, there are a plethora of experimental methods for designing and selecting antibody therapeutics. Despite advances in the field of experimental antibody development, computational antibody analysis and design remains a relatively emergent field. If successful, an *in-silico* antibody design platform could alleviate months of expensive experimental work. In addition, the computational methods would have an ability to target a specific epitope of a protein, which is challenging to accomplish in an experimental setting. Furthermore, a successful pipeline for computational antibody development would offer substantial proof that, as a community, we understand antibody functionality and the factors that drive it.

In this chapter, we describe a novel computational method for antibody design. Our aim is to create an algorithm that can design high-affinity, human binders, targeting a specific

epitope of a protein. Previous computational antibody design work (Pantazes et al. 2010; Li et al. 2014; Lapidoth et al. 2015) has tended to assume that CDR structures can be combined without any restrictions and grafted onto any framework. But decades of experience of humanization studies indicate this is not the case – while the antibody structure is extraordinarily malleable to structural modifications, transplanting the CDRs onto a non-native framework can often lead to a drastic loss of affinity (Queen et al. 1989). In any design pipeline, errors tend to propagate forward, reducing the accuracy. Therefore, it is crucial to reduce the errors at the earliest steps in the methodology. We set out to remove early errors by using a large dataset of antibody sequences as a foundation on which we build our design pipeline. By using a set of human sequences, we also improve the chances that our designs will not cause an immunogenic response.

At any given point in time, a healthy human immune system has ~10^9 distinct B-cell receptor sequences available. Some of that diversity can be captured using the modern, high-throughput Next-Generation Sequencing (NGS) techniques. In this chapter, we used an antibody NGS dataset containing over 5,000,000 sequences of each chain type (κ, λ, heavy) to create a computational antibody design algorithm. First, we reduced the initial sequence volume into a library of ~20,000 three-dimensional, structurally diverse antibody models. To create this library, we used the fast VH-VL orientation assignment algorithm developed in Chapter 3 and knowledge-based structural characterisation of CDRs. Next, we created a method to computationally pan a selected epitope of an antigen against the model library. This was accomplished using ZDOCK (Pierce et al. 2011), a Fast-Fourier Transform docking protocol. Every model was docked to the selected antigen and the top scoring poses were selected using ranking method based on the LambdaMART algorithm (Burges 2010). Finally, we used the Rosetta scripts

(Fleishman et al. 2011) framework to build a custom computational affinity maturation script that simulated the somatic hypermutation phase of natural antibody development. We tested our computational design pipeline by creating antibody binders against the Hen-egg lysozyme. We show that the binders we created have similar properties to the real antibody crystal structures, available in the SAbDab database (Dunbar et al. 2016). Finally, we show that the binders we built have predicted immunogenicity similar to natural human antibody sequences, increasing the chances that they could be easily developed into therapeutics.

5.2 Methods

5.2.1 Next-Generation Sequencing dataset

Our computational antibody design protocol starts by deriving a library of models of human antibodies from a dataset of Next-Generation Sequencing (NGS) results. To design this library, we have used an NGS set provided by our collaborators at UCB Pharma Ltd. This dataset has been described previously in Chapters 3 and 4. Briefly, the set contains around 5,000,000 sequences of unpaired IgM chains of each type (κ, λ, Heavy) derived from ~500 individuals, mostly of Asian origin. The exact numbers are shown in Table 3.2.

The NGS sequences were numbered according to both the IMGT scheme (Lefranc et al. 2003) and Chothia scheme (Al-Lazikani et al. 1997) using the ANARCI software (Dunbar et al. 2015). The ANARCI software aligns the sequence to be numbered to Hidden Markov Models (HMMs) constructed from sequences with known numbering. The CDR definitions for the two numbering schemes are shown in Table 5.1.

CDR type	IMGT CDR definition	Chothia CDR definition
CDR-L1	27 – 38	24 – 34
CDR-L2	56 – 65	50 – 56
CDR-L3	105 - 117	89 – 97
CDR-H1	27 – 38	26 – 32
CDR-H2	56 – 65	52 – 56
CDR-H3	105 – 117	95 – 102

Table 5.1 CDR definitions. The table shows the two CDR definitions used in this chapter. The definitions are shown in their original numbering scheme.

Before this raw dataset was used to construct the model library, we filtered the constituent sequences based on our ability to model their structural elements.

5.2.2 High-throughput structural assignment of CDRs and frameworks

In order to ensure that high quality models would be built, we filtered out sequences from our NGS dataset where not every CDR could be modelled. First, we extracted all Chothia defined CDR sequences from the dataset and filtered out the repeating sequences (see Table 5.2).

	CDR-L1	CDR-L2	CDR-L3	CDR-H1	CDR-H2	CDR-H3
Unique	1,042,157	162,429	1,099,564	116,141	164,106	1,688,049
Redundant	9,371,465	9,371,465	9,357,354	5,645,307	5,645,307	5,630,528

Table 5.2 The number of Chothia defined CDR sequences of each type in our NGS dataset. The "Unique" row shows the number of sequences that are different by at least one amino acid. The "Redundant" row shows the total number of sequences. The reason for the differences between redundant numbers for CDRs on the same chain is that some sequences are missing their CDR3 region.

The ability to model the CDR sequences was assessed by Cristian Regep using the knowledge-based loop modelling software FREAD (Choi et al. 2010). FREAD takes as input the sequence of the loop to be modelled (the target) and the protein structure, on which the loop should be grafted (the framework). The possible loop models are then recalled from a template database using the following considerations. First, the Environment Specific Substitution Table (ESST) scores are calculated for each template in the database. The ESST scores are taken from a table, which quantifies how likely it is to substitute the template amino acid with the target amino acid, given the dihedral angles of the template residue. After the suitable templates have been identified, using an ESST score cut-off, the algorithm considers the difference in orientation and backbone RMSD between the loop anchors on the framework structure and the original anchors of the template. Finally, FREAD checks if there are any clashes between the model and the framework. If at any step no available templates are identified (e.g. no templates above a ESST cut-off, or above anchor RMSD cut-off) the algorithm reports that the target cannot be modelled using the given template database. This formula has the disadvantage that not all sequences can be modelled, but puts high confidence in the

model when the algorithm succeeds. FREAD has been shown to outperform the more complex ab initio methods, where suitable templates were available (Choi et al. 2010).

To structurally characterize the NGS sequences, first the framework structures were assigned to each antibody chain sequence. Towards this purpose, the Chothia defined frameworks of the NGS dataset were compared to the sequences of all the frameworks of known structures. These structures were collected from the SAbDab database (Dunbar et al. 2014). The highest sequence similarity framework was assigned to each NGS sequence. Next, a database of CDR structures was constructed by extracting the Chothia defined CDR structures from all the antibodies available in the SAbDab database. The ESST cut-offs used for each loop length and CDR type were set such that 80% of CDRs would be modelled below 1.0Å. CDR shapes were then assigned to each NGS sequence using ESST scores calculated between the sequences of the NGS loops and the structure of the model loops, and the anchor orientation between the CDR model and the assigned structural framework. Any NGS sequence where we could not model all 3 CDRs was filtered out.

Due to a lack of data, modelling CDR-H3 loops longer than 15 residues was unreliable. NGS sequences containing these longer CDR-H3s were removed from our dataset. This constraint had a separate effect of ensuring that our antibody models all had a flat or concave binding site, improving the chances of contacting the antigen with multiple CDR loops and, therefore, increasing the number of residues available for modification.

After filtering the NGS data by our ability to model the CDRs of the NGS sequences we greedily clustered the remaining set.

5.2.3 Greedy clustering of sequences to identify a diverse set

To be able to bind a wide range of antigen structures, our Antibody Model Library (AML) should contain a diverse set of binding site shapes. To increase the structural diversity within our set we greedily clustered the remaining NGS sequences.

To cluster the remaining NGS sequences we used the CDHIT (Fu et al. 2012) software. This algorithm first sorts the sequences in order of decreasing length. The sequences are then clustered using a greedy approach, by applying the following steps. The first sequence in the list becomes the representative of the first cluster. After that, the sequence identity is calculated between each following sequence and every cluster representative. The sequence is matched to a cluster for which the sequence identity is above the pre-defined threshold. If none of the comparisons results in an identity above the threshold, the sequence becomes a representative of a new cluster. This procedure is followed until all sequences have been classified. The program uses a "short-word filtering" algorithm for fast sequence comparison. This short-word filtering utilizes the fact that the sequences must share a certain number of short polypeptide substrings, called "words", for the identity to be above the threshold. The occurrence of each such word is calculated for both sequences to be compared. As soon as it clear that the number of identical words fails to meet the required threshold, the sequence pair is classified as dissimilar.

We clustered the fully modellable antibody sequences using CDHIT with a 90% sequence identity threshold, with a condition that all the sequences in any given cluster must be of the same length. Since most of the length differences can be attributed to the CDRs, this additional condition prevented discarding potentially structurally distinct sequences. We only retained clusters containing more than 10 sequences (Friedensohn et al. 2017), to

reduce the error present in our dataset. From each remaining cluster, we picked the sequence with lowest median sequence identity to other cluster members ("middle" sequence) for further analysis.

5.2.4 Creating complete variable regions

The NGS dataset contains only single-chain antibody sequences, without any heavy-light chain pairing information. Even after the CDHIT clustering, the size of the dataset allowed for over 20,000,000,000 potential heavy-light chain pairs. To create an AML containing models of full Fv sequences we needed a method to efficiently select pairs of antibody chain sequences which are likely to form viable antibodies and model the orientation between the chains. Towards this goal, we employed the high-throughput method described in Chapter 3, which involves rapid calculations of sequence identity over the interface residues.

The method involves identifying interface residues by calculating the buried solvent area for each residue contained within a set of 989 Fv structures from SAbDab (Dunbar et al. 2014) database. The interface sequence identity is then calculated by comparing the interface residues from the real antibody structures and the predicted interface residues within the NGS sequences. If the identity to any structure is over a designated threshold (0.82 in our study), the orientation of that structure is assigned to the Fv sequence. If more than one structure fulfils the threshold requirement, the orientation of the structure with highest sequence identity is selected. By expressing the sequence identity calculation using sparse matrix multiplication our method can quickly analyse billions of interfaces. The Fv sequences failing to meet the 0.82 interface sequence identity threshold were discarded, reducing the size of the set of potential Fv sequences by three orders of magnitude.

At this point in our study, the selected Fv sequences had VH-VL orientation and CDR structural models assigned to them. We used those structural assignments to filter the remaining sequences by their predicted binding site shape diversity.

5.2.5 Selecting a structurally diverse set

As described above, we used FREAD (Choi et al. 2010) to assign the CDR models and interface residue identity calculations to assign VH-VL orientations. Those assignments were next used to greedily cluster the remaining Fv sequences, based on their estimated structural diversity.

First, we calculated distance matrices between all CDR models that were used to construct the FREAD database (see methods Section 5.2.2). We have done this in a length-independent fashion, using the Dynamic Time Warping (DTW) method outlined in Chapter 2. Next, we calculated the orientation RMSD distance matrix between the 989 orientation templates (see Chapter 3). This allowed us to pre-calculate structural comparisons between each constituent part of our structurally characterised paired NGS sequences and reduce the structural comparison between the binding sites of these Fv sequences to distance matrix lookups.

At this stage, each one of our remaining Fv sequences had six CDR templates and one orientation template assigned to it. We checked if two Fv sequences code for structurally distinct binding sites using the following steps:

1. We looked up the orientation RMSD between the orientation templates. If the RMSD is above our identicality threshold of 1.5 Å (see Chapter 3), we classified the binding sites as distinct without considering the CDRs.

2. If the orientation RMSD was below 1.5 Å, we calculated the CDR distance score using the following formula:

$$D_{CDR} = \sqrt{\frac{\sum_{X}^{(L1-L3,H1-H3)} {DTW_{CDR-X}}^2 * \max(L_{CDR-X,1}, L_{CDR-X,2})}{\sum_{X}^{(L1-L3,H1-H3)} \max(L_{CDR-X,1}, L_{CDR-X,2})}}$$

Where DTW_{CDR-X} is the DTW distance between the CDR templates, resulting from a lookup in the distance matrix, $\max(L_{CDR-X,1}, L_{CDR-X,2})$ is the maximum length of the two loops and the summation is over the CDR types.

3. We checked if the D_{CDR} is above a pre-defined threshold $T = 1.0\text{Å}$. If it was, we classified the two Fv sequences as structurally distinct. In practice, we compared ${D_{CDR}}^2$ to T^2 to reduce the computational time needed to calculate the square root.

Using the above comparisons, we clustered the remaining Fv sequences. To perform the clustering, we followed a greedy clustering approach, similar to the one implemented in the CDHIT software (see Section 5.2.3), except that an Fv sequence would have to pass both the orientation RMSD threshold and the CDR similarity threshold to be classified into a given cluster.

5.2.6 High-resolution modelling and clustering of Fv models

To model our structurally diverse Fv sequences we used the ABodyBuilder software (Leem et al. 2016). This is a high-resolution modelling method designed to predict three-dimensional coordinates for all atoms within the Fv sequence. First, the software numbers the sequences according to the IMGT scheme (Lefranc et al. 2003) and separates the sequence into a framework part and the constituent CDRs. Next, the software selects a framework model from a database of antibody crystal structures using sequence identity. Then, the VH-VL orientation is predicted, using the ABangle

parameters (Dunbar et al. 2013). Following that, the CDR structures are predicted using the FREAD methodology (Choi et al. 2010). Finally, the side chain orientation is modelled using SCWRL4 (Krivov et al. 2009). When benchmarked on targets from the second Antibody Modelling Assessment (Teplyakov, Luo, et al. 2014), ABodyBuilder performed similar to the state-of-the art pipelines, but at a significantly reduced computational cost (Leem et al. 2016). Using ABodyBuilder we could model the remaining ~66,000 Fv sequences in 19 CPU hours.

The models created by ABodyBuilder were then clustered again, using the full three-dimensional coordinates. First, for each pair of models, the common CDR residues, present in both structures, were identified using IMGT numbering. The models were then aligned, in pairs, using the Kabsch algorithm (Kabsch 1976) and the RMSD was calculated over the common CDR residues. The resulting distance matrix was then clustered using the Unweighted Pair Group Method with Arithmetic Mean (UPGMA) hierarchical clustering method (Sokal et al. 1958). We chose a clustering cut off 1.0Å, to provide us with a structurally diverse set of binding sites. The selected AML models were then minimized using the Rosetta protocol (Conway et al. 2014).

5.2.7 Docking and pose selection

In order to test our AML, we computationally panned the constituent models against known epitopes of Hen-egg lysozyme. To identify the epitope residues of a given lysozyme configuration in an Antibody-Antigen complex, we calculated differences in Solvent Accessible Surface Area (SASA), with antibody present and absent from the complex. To perform the SASA calculations we used the NACCESS software (Hubbard et al. 1993). NACCESS calculates the accessible surface area by rolling a probe over the Van der Waals surface of a protein structure. For our calculations, we used the default

probe size of 1.4 Å. We selected epitope residues as any antigen residue that had a dSASA greater than 0.

To pan a given antigen against the AML we used the ZDOCK software (Pierce et al. 2011). ZDOCK represents the antibody and the antigen using grids and explores possible docking poses using a three-dimensional Fast Fourier Transform (FFT). The use of FFT greatly improves the computational efficiency of the algorithm, allowing it to probe thousands of possible docking poses in minutes. We chose ZDOCK as it has been shown to be able to efficiently identify the CDRs and the neighbouring residues as the most likely binding site of an antibody (Mintseris et al. 2007). The poses are then scored using a function that includes electrostatics, shape complementarity, and a statistical potential. We built 1,000 docking poses per each AML model, resulting in 19,019,000 docking poses per antigen epitope.

Using the ZDOCK scoring function we selected the five highest scoring antibody-antigen poses per model for further analysis. These top scoring poses were further optimized using the high-resolution phase of Rosetta docking (Chaudhury et al. 2011) algorithm. During this high-resolution refinement, the following steps were repeated in a Monte-Carlo fashion (Chaudhury et al. 2011):

1. The ligand coordinates were perturbed by a random direction and magnitude sampled from a normal distribution, centred around 0.1 Å and 3.0˚
2. The antigen orientation was energy-minimized
3. The side-chain orientations were optimized

These three steps are repeated 50 times, with the new orientation being accepted or rejected based on the Rosetta energy (Chaudhury et al. 2011). During this high-

resolution phase, the residues are represented in an all-atom fashion (Chaudhury et al. 2011).

These optimized poses were then ranked using several quality scores, calculated by Rosetta, the original ZDOCK score and a custom statistical potential, calculated using available antibody-antigen complex structures (see below). The Rosetta scores were calculated using the InterfaceAnalyzer application (Stranges et al. 2013). We calculated the following scores: Buried surface area, Surface Complementarity, Packstat (Sheffler et al. 2009) and predicted affinity. The Surface Complementarity and Packstat are scores expressed by a value ranging from 0.0 – 1.0, reflecting how the docking pose compares to real protein-protein complexes (0.0 no similarity, 1.0 perfect similarity). The predicted affinity was calculated as a difference in Rosetta energy between the complex, and separated antibody and antigen.

To gather statistics about existing antibody-antigen complexes and to design our own custom statistical potential (see below) we created a set of structurally diverse antibody-antigen complex structures. First, we extracted 1,370 complex structures from the SAbDab database (Dunbar et al. 2014), ensuring that each extracted complex had all their CDRs crystallographically resolved. Next, we calculated MM-scores (Mukherjee et al. 2009) between each pair of complexes, and clustered the resulting distance matrix using UPGMA algorithm at 0.5 MM-score cut-off. This resulted in a set of 280 structurally-distinct antibody-antigen complexes.

The custom statistical potential we used was constructed using the following steps:

1. We calculated Cβ – Cβ distances for every interface residue pair within each complex in our set of 280 structurally-distinct antibody-antigen complexes (see above). We ignored Glycines in this calculation.

2. We identified interacting residue pairs using the Arpeggio algorithm (Jubb et al. 2017) and calculated the maximum and minimum Cβ – Cβ distances for each amino acid pair. These distances defined an interacting distance bracket.
3. We counted the number of amino acid pairs within interacting distance for each complex within our dataset.
4. Using ZDOCK we created a set of 1,000 "reference" docking poses for each of the 280 antibody-antigen complexes within our set, restricting the epitope and paratope residues to the ones observed in the original complex.
5. We repeated the steps 2 and 4 for these reference poses.

The statistical potential for each amino acid pair was then calculated using the following formula (Zhou et al. 2002):

$$U(aa_1, aa_2, \Delta r) = -RT \ln \frac{1000\, N_{obs}(aa_1, aa_2, \Delta r)}{N_{ref}(aa_1, aa_2, \Delta r)}$$

Where aa_1 and aa_2 are the amino acids in the pair (e.g. Arginine – Lysine), R is the gas constant, T is the room temperature (300 K), $N_{obs}(aa_1, aa_2, \Delta r)$ is the count of aa_1 and aa_2 amino acid pairs within Δr interacting distance bracket in the set of true antibody-antigen complexes and $N_{ref}(aa_1, aa_2, \Delta r)$ is the corresponding count within the reference poses. The factor of 1000 is introduced in the fraction to account for the number of reference docking poses per each antibody-antigen complex.

The quality scores described above (ZDOCK score, Buried surface area, Surface Complementarity, Packstat, predicted affinity and statistical potential score), were combined into a ranking function using the LambdaMART algorithm (Burges 2010). The algorithm learns to replicate a ground-truth ranking, using a set of features calculated for each object to be ranked and is a modification of the RankNet (Burges 2010) method,

used by Microsoft in the Bing search engine. The objective of the RankNet method is to minimize the number of inversions in the ranking, where an inversion is an incorrect ordering among a pair of results (Burges 2010). LambdaMART improves upon the RankNet method by optimizing the learning process and using gradient boosted decision trees for prediction. The ranking quality, optimized by the algorithm is given by the Normalized Discounted Cumulative Gain (NDCG) (Y. Wang et al. 2013). LambdaMART was the winning submission in the Yahoo! Learning to Rank Challenge (Chapelle 2011) and performed well in other large-scale tests of machine learning ranking algorithms (Tax et al. 2015).

The ground-truth ranking was set to quantify similarities between poses and the true antibody-antigen complex. The objective ranking score was calculated by comparing three types of electrostatic interactions (Ionic, Hydrogen Bonds and Van der Walls) between the pose and the true complex. We only counted the interactions that were of the same type and to the same antigen residue (e.g. an Ionic interaction with Lysine-96). The interactions were identified using the Arpeggio software (Jubb et al. 2017). The score was calculated through a weighted sum of fractions of preserved contacts of each type. The weights were 4 for ionic interaction, 2 for hydrogen bond and 1 for Van der Walls interaction, reflecting the relative strength of these interactions.

The ranking method was trained and tested in a take-one-out fashion on a group of four sets of docking poses, created for the Lysozyme epitopes identified within our set of 280 structurally diverse antibody-antigen interfaces. The algorithm was trained on three sets of poses and tested on the fourth.

5.2.8 Computational affinity maturation

The 100 highest-scoring docking poses. given by the selection procedure described in Section 5.2.7, were selected for sequence modification through computational affinity maturation.

We constrained the available substitutions at the antibody paratope by creating Position Specific Substitution Tables (PSSTs), separately for each CDR sequence within our library. First, we have extracted CDR sequences from our original NGS dataset, separately for each CDR type-CDR length-Variable germline gene combination (e.g. CDR-L1, length 11, IGKV3-20). Next, for each position within each pool of CDR sequences, we selected a set of allowable substitutions. We only allowed substitutions to an amino acid, if the prevalence of that amino acid for a particular position was above 2%. This constraint was set to eliminate substitutions possibly originating from experimental sequencing errors (Galson et al. 2016). The non-CDR positions were forced to remain fixed during the maturation process.

To improve the predicted affinity of the antibody towards the antigen we developed a Rosetta Scripts protocol (Fleishman et al. 2011). The first step selected interface residues using logic created by Stranges & Leaver-Fay for their computational protein design workflow (Leaver-Fay, Jacak, et al. 2011). This method selected residues according to two criterions. The first criterion accepted residue pairs at antibody-antigen interface which contained atoms within 5.5Å of each other. The second criterion selected residue pairs whose $C_1\alpha$ - $C_1\beta$ - $C_2\alpha$ angles were within 75° (the subscript in the formula indicates which residue the atom belongs to) and contained atoms within 9Å threshold. This restrictive selection logic ensures that only residues which have a potential of interacting, or are already interacting, are selected. The second step in the

mutation process involved randomly selecting one of these interface residues for substitution by a different, randomly selected, amino acid included in the relevant PSST. The third step perturbed the backbone of both the antibody and the antigen using a backrub move (Smith et al. 2008). This perturbation allowed for some flexibility to be included in the design process and has been shown to improve design outcomes (Babor et al. 2011). Finally, the backbone and side-chains of the complex were minimized using the Rosetta energy. After these steps, the predicted change in affinity was calculated. This change is used as a criterion for accepting or rejecting the mutation. In addition to the aforementioned steps, which were repeated in a Simulated Annealing trajectory, every 50 steps the docking pose was locally optimized, using the high-resolution phase of Rosetta's docking protocol (Chaudhury et al. 2011). We followed the trajectory for a total of 2,000 steps.

5.2.9 Immunogenicity analysis

One of the metrics we used to analyse the matured poses was the immunogenicity score, calculated using the methodology described by King et al., 2014 The score consists of three parts. The first part is based on a SVM (Support Vector Machine) score calculated over each 15-residue long window, which attempts to predict how likely the corresponding peptide sequence is to bind a Major Histocompatibility Complex (MHC). This SVM was trained on immunological data from the Immune Epitope Database (IEDB). The second part of the score is calculated on each 9-residue window and compares the frequency of the 9-mer in the host genomic data and in the known epitope data (a sequence occurring with a high frequency in a human genome would be rewarded, while the opposite is true for sequences occurring in the known epitope data). The third part penalizes any deviations from the original charge of the protein. These three parts were

combined with a standard Rosetta score that measures the stability of the protein. The weights assigned to each segment were calibrated on existing protein structures.

5.3 Results & discussion

5.3.1 Overview of the design pipeline

Our design strategy started with the construction of a human, structurally diverse Antibody Model Library (AML). These models were derived from a large, human Next-Generation Sequencing (NGS) dataset of human IgM sequences, containing ~5,000,000 unpaired, full chain sequences of each type (κ, λ, heavy). Using high-throughput structural characterization techniques, pairs of heavy-light chain sequences were selected from this initial pool of data which were likely to be modelled with high accuracy. The modellable heavy-light sequence pairs were then filtered based on the predicted VH-VL orientation RMSD and CDR structural similarity. Next, high-resolution Fv models, were created using the ABodybuilder software (Leem et al. 2016). Finally, the structural diversity of the dataset was enriched by removing highly homologous models. The entire filtering procedure reduced the NGS dataset to an AML containing 19,019 Fv models which were then used as a base for docking and pose optimization. The library construction workflow is summarized in Figure 5.1.

Once the AML had been created, our method created potential binding poses between the antibody models and a selected epitope on an antigen. The pipeline took as input the crystal structure of the antigen and selected epitope residues. Using the ZDOCK software (Pierce et al. 2011) we created 1,000 binding poses for each model-antigen pair, resulting in a set of 19,019,000 potential poses. The top 5 highest-ranked poses from each model-antigen pair (as measured by the ZDOCK energy function (Pierce et al. 2011)) were optimized using the high-resolution docking protocol implemented in the

Rosetta software (Chaudhury et al. 2011) resulting in a set of 95,095 binding models. From this set, only the poses containing a large fraction of epitope and CDR residues within close proximity to each other were retained. These poses were then scored using the InterfaceAnalyzer routine from Rosetta (Stranges et al. 2013). The resulting scores, along with the ZDOCK potential and a custom antibody-specific statistical potential, were used as input to a machine learning algorithm that ranked them according to their similarity to available crystal structures of antibody-antigen complexes. The highest ranked 100 poses were then passed on to the affinity maturation phase.

To increase the predicted affinity of the Fv models towards the antigen, Position Specific Substitution Tables (PSST) were used. These were built from our NGS data, separately for each (Germline, CDR type, CDR length) combination. Restricting the sequence modifications to naturally occurring residues should ensure that the matured sequences remain close to the human antibody repertoire. To redesign the Fv surface, we used a Rosetta scripts (Fleishman et al. 2011) protocol. At each time step, the script introduced a random allowed mutation to one of the residues in the antibody binding site. The resulting mutant was then minimized and the ΔΔG of binding was estimated. These steps were repeated in a Simulated Annealing trajectory for 2,000 steps, with an additional redocking every 50 steps. Four trajectories were calculated for each one of the top 100 poses resulting in a pool of 400 final poses from which the top binders were selected based on manual inspection and predicted affinity.

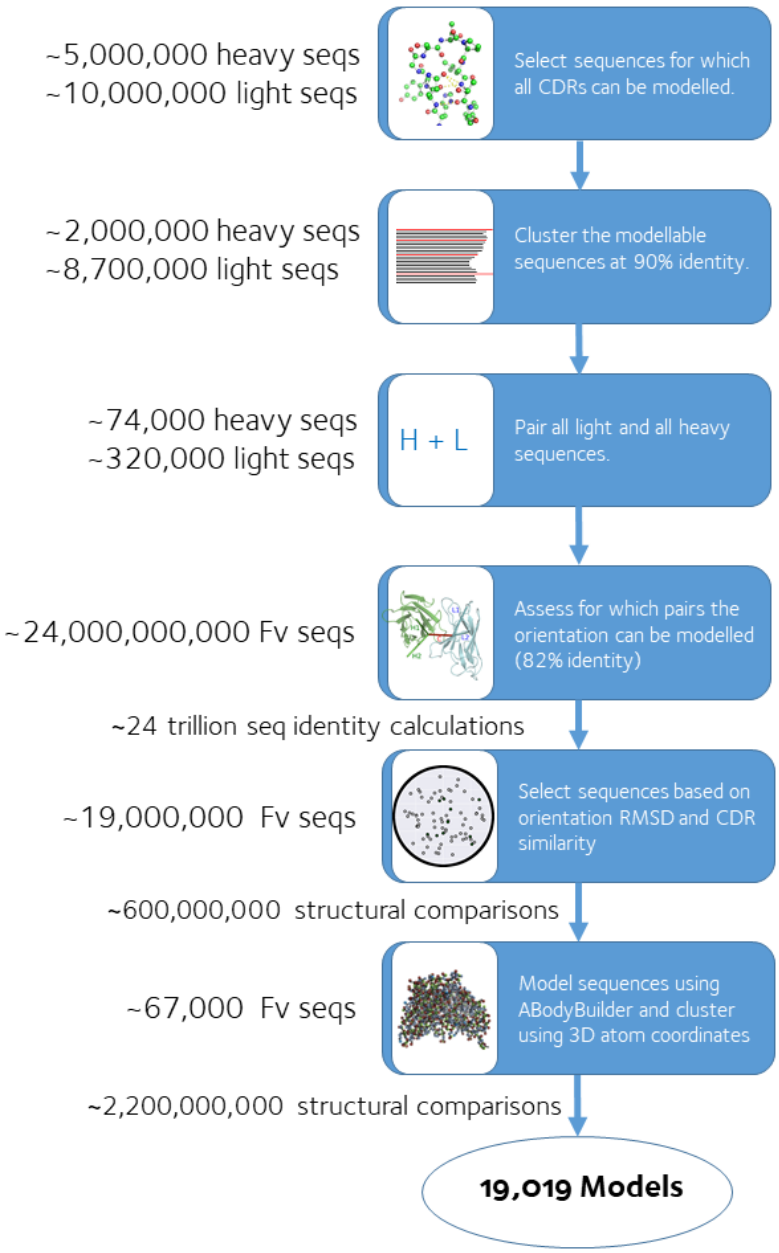

Figure 5.1 Antibody Model Library (AML) construction process. The flowchart summarizes the steps involved in the construction of the AML, used in the design process. The numbers on the left of the flowchart show the approximate number of sequences at each stage of the workflow and the number of operations involved at the final steps. The pipeline starts with a large, NGS dataset of antibody chain sequences containing ~5,000,000 sequences of each chain type and reduces this set to an AML of ~20,000 high-quality models. The process can be readily adapted to work with any other NGS set. The entire process takes about two weeks on a 70-core cluster.

5.3.2 NGS sequence clustering and selection

The library of Fv models was constructed from an NGS dataset containing ~15,000,000 sequences of human IgM chains obtained from around 500 individuals. We used the protocol described in Section 5.2.3 to select a diverse set of sequences, which could be modelled with high accuracy, from the initial pool of data.

First, the ability to model the CDR sequences was analysed. This was done by Cristian Regep using the knowledge-based software FREAD (Choi et al. 2010). Only sequences where all three CDRs could be modelled accurately were retained. This step reduced the size of our set to 5,035,560 κ chains, 3,716,485 λ chains and 2,078,859 heavy chains. As expected, the number of rejected heavy chain sequences was highest, due to the difficulty of modelling the CDR-H3 loop (Weitzner et al. 2015).

Next, we used the greedy sequence clustering program CD-HIT (Fu et al. 2012) to reduce the redundancy within our set. We clustered the sequences at 90% amino acid sequence identity, forcing all clusters to contain sequences of the same length. We only retained clusters if they contained at least 10 sequences. This step reduced the size of our set to 153,646 κ chain sequences, 169,020 λ sequences and 74,320 heavy chain sequences.

Following the greedy sequence clustering we investigated pairing between chains. The pairing preferences of different germlines is an area of active research (Jayaram et al. 2012), therefore we focused on our ability to accurately predict the orientation between a given antibody heavy and light chain. From our analysis (see Chapter 3) we find that we can accurately assign the VH-VL orientation for sequence pairs with interface sequence identity above 0.82 to an Fv which has been structurally characterized. We paired all remaining light and heavy chain sequences in our set

(23,980,537,120 pairs) and compared their interface sequence to that of the 989 Fv structures in our template set (see Chapter 3). This procedure resulted in ~2.35 x 10^{13} interface comparisons. Keeping only those pairs with interface sequence identity above 0.82 to a known structure reduced the size of our dataset to 1,217,860 heavy-λ sequence pairs and 17,919,229 heavy-κ sequence pairs. The disparity between the number of λ and κ sequences is caused by the overrepresentation of κ chains in the structural data (Dunbar et al. 2014).

The remaining Fv sequences were then further filtered using high-throughput structural characterization methods to obtain a model library containing a diverse range of binding site shapes.

5.3.3 Structural characterization and clustering

To further reduce the size of the model library we performed greedy clustering on the remaining 19,137,089 Fv sequences, based on their structural diversity.

First, we assigned VH-VL orientation and CDR structural models to each Fv sequence, based on their closest interface sequence identity partner and the previously-calculated FREAD predictions respectively. Then, we clustered the sequences using the methodology described in Section 5.2.5. This clustering step ensured that the model library would contain a diverse set of binding site shapes. Our high-throughput structural clustering procedure reduced the library size to 66,783 Fv sequences.

In the next stage, these ~66,000 Fv sequences were modelled using the ABodyBuilder (Leem et al. 2016) software. The constructed models were then clustered again, using RMSD distance between the equivalent CDR residues and the hierarchical clustering algorithm (Ward 1963) at 1.0Å threshold (see Section 5.2.6). This final modelling and

clustering steps reduced the library size to 19,019 models. The models were then relaxed using Rosetta (Conway et al. 2014) protocol.

5.3.4 Properties of the antibody models within Antibody Model Library

We compared the models in the AML to the crystal structures of 64 antibody therapeutics, available in SAbDab database (Dunbar et al. 2014) (see Figure 5.2). The models were aligned with the therapeutics and the TM-scores (Zhang et al. 2004) were calculated. We used TM-score here instead of RMSD, because the antibodies in our set span a range of sequence lengths (see below) and the TM-score allows for length-independent structural comparisons (in contrast to RMSD which is length-dependent). We find that for all therapeutics we can find at least one model in our library with a TM-score of over 0.90 and for 57 out of 63 we can find a model with a TM-score of over 0.95. This suggests that the AML has excellent structural coverage of currently structurally characterised therapeutics.

We compared the number of CDR residues (as defined by IMGT numbering (Lefranc et al. 2003)) in the AML and a set of 137 antibody binders (Burkovitz et al. 2016). We found that the values are very similar with average combined CDR length of 48.07 for therapeutic antibodies and 48.80 for NGS models (see Figure 5.3). This result shows that the models within our library contain, on average, a similar number of CDR residues as the therapeutic antibodies.

In addition to containing similar structural features to the true therapeutic antibodies, our model library has been constructed from purely human sequences, which should significantly reduce the immunogenicity of our designs (see Section 5.3.7).

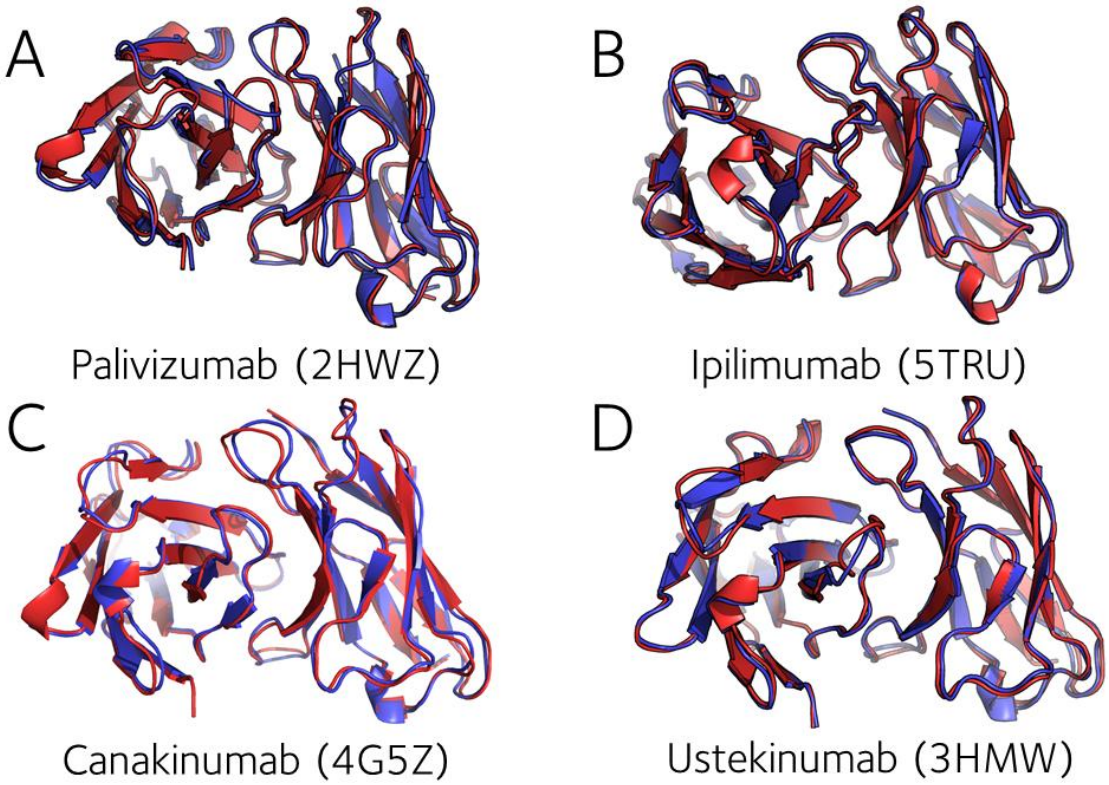

Figure 5.2 Four therapeutics and their closest structural matches in the AML. The Figure shows alignments between crystal structures of four therapeutic antibodies and the structurally closest NGS models within our library (as measured by TM-score). In each panel the crystal structure of the therapeutic is shown in red and our model is shown in blue. A) The crystal structure of Palivizumab (Synagis) aligned with its closest structural homologue (TM-score 0.951). B) The crystal structure of Ipilimumab (Yervoy) aligned with its closest structural homologue (TM-score 0.986). C) The crystal structure of Canakinumab (Ilaris) aligned with its closest structural homologue (TM-score 0.987). D) The crystal structure of Ustekinumab (Stelara) aligned with its closest structural homologue (TM-score 0.990). The figure shows that our library contains enough structural diversity to contain close structural homologues for most known antibody therapeutics.

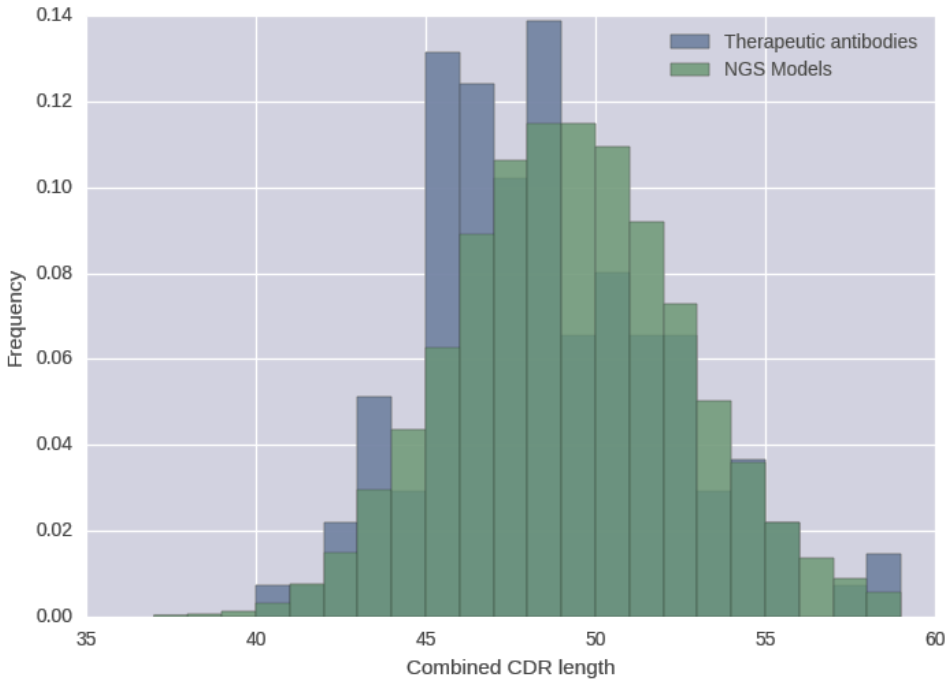

Figure 5.3 The comparison of the combined CDR length between the Therapeutic antibodies and the NGS Models. The figure shows the total number of CDR residues (selected using IMGT definition (Lefranc et al. 2003)). The average total length was 48.07 for therapeutic antibodies and 48.80 for NGS models. The figure shows that the NGS models we built contain similar range of CDR lengths as therapeutic antibodies.

It is important to note here that the antibody structural data available in the PDB is most likely not statistically representative of the human antibody structural space (e.g. Dunbar et al. 2014; Weitzner et al. 2015). As the crystallography experiments are costly, experimentalists typically focus on observations which are therapeutically relevant (e.g. Hu et al. 2013) or have unusual sequence (e.g. Scharf et al. 2014). A high number of structurally characterised antibodies have also been heavily engineered, for example through CDR grafting or chimerisation (e.g. Teplyakov et al. 2010). There is also an overrepresentation of antibodies with kappa chains, as most therapeutically relevant antibodies utilize this light chain type (Nowak et al. 2016). There is also a significant bias towards lysozyme binders (~6% of antibody-antigen complex structures contain

lysozyme as the antigen) as this antigen is typically used as the "standard" target for studying antibody binding modes (Dunbar et al. 2014). Furthermore, protein crystallization experiments are prone to failure (Grey et al. 2010), which likely further biases the PDB against antibodies that do not readily crystallize. This is also the hardest bias to quantify, as failed experiments are usually not reported. As our model library has been constructed using algorithms trained and validated using structural data from the PDB, it is likely to replicate the biases present in the database.

It is difficult to comment on the theoretical size of the structural space available to the antibodies. Nevertheless, as the CDR-H3 structural diversity is considered to be much higher than that of other CDRs (Weitzner et al. 2015) we can anticipate that the size of that space is likely much larger than the portion captured by our model library of ~19,000 antibodies. The diversity of the human antibody space is likely necessary to create binding sites complementary to most possible protein surfaces (Chaffey et al. 2003), as any gaps in recognition are likely to be exploited by opportunistic invaders. It is unlikely that our AML contains binding sites complementary to every possible protein surface. Nevertheless, we showed in this section that the diversity represented in our library is high enough to contain binding sites similar to those of the experimentally characterized antibody therapeutics, indicating that our AML could potentially be useful for developing therapeutics against some pharmaceutically relevant targets.

In the next section, we describe how the AML was used to computationally design binders against selected epitopes of target antigens through docking and computational affinity maturation.

5.3.5 Docking

We have calibrated our docking pipeline on four distinct antibody-lysozyme complexes with available crystal structure (PDB (Berman et al. 2000) ids 3D9A, 1VFB, 4TTD, 1P2C). These complexes were selected by clustering all available antibody-antigen complex crystal structures (see Section 5.2.7) and identifying four containing a distinct lysozyme epitope. The lysozyme structures were extracted from the complexes and relaxed using Rosetta (Conway et al. 2014) to remove any information about side-chain conformations. The Lysozyme epitopes were identified for each complex using SASA calculations (see Section 5.2.7).

We docked our model library against each of the four lysozyme epitopes using the program ZDOCK 3.0.2 (Pierce et al. 2011) restricting the epitope residues to the ones identified for each complex. We created 1,000 poses for each model resulting in 19,019,000 docking poses per lysozyme epitope (76,076,000 poses in total). From this set we selected the top five scoring poses for each antibody model using the ZDOCK ranking. The selected 95,095 poses per lysozyme epitope (380,380 poses in total) were then minimized using the Rosetta high-resolution docking phase (Chaudhury et al. 2011). This step was performed to remove any structural clashes and prepare the docking poses for calculation of features used in our pose ranking protocol. To ensure good coverage of the selected epitope, and maximize the number of designable residues, poses which contained less than 65% of selected epitope residues or less than 40% of the CDR residues at the interface were removed from the set. For a residue to be considered part of the interface, its Cα must be within 12Å of the Cα of the closest residue on the binding partner. The number of poses passing through this filter, for each epitope we tested, is shown in Table 5.3.

Our next objective was to create a method that would reduce the pool of poses (see Table 5.3) to a smaller set of 100 models that could be further designed through computational affinity maturation. To decide on good poses, we compared the properties of a pose with those found in antibody-antigen complexes with a known crystal structure. In particular, a good pose should resemble the available crystal coordinates in terms of types of contacts formed. To perform this selection, a custom machine learning protocol was built with the objective of ranking the poses according to their similarity to known antibody-antigen complexes (see methods Section 5.2.7).

PDB id containing the target epitope	Number of docking poses selected
3D9A	10,786 (11%)
1VFB	21,490 (23%)
4TTD	7,377 (8%)
1P2C	25,629 (27%)

Table 5.3 Number of docking poses selected for further analysis. This table shows how many poses remained from the original set of 95,095 after applying the paratope-epitope filter, requiring the poses to contain at least 65% selected epitope residues and at least 40% CDR residues in the interface. For a residue to be considered part of the interface, its Cα must be within 12Å of the Cα of the closest residue on the binding partner. The PDB id column specifies which crystal structure has the respective lysozyme epitope been taken from.

Our pose ranking protocol is based on the LambdaMART ranking algorithm (Burges 2010). The feature set was composed of six elements. The first four were calculated using the Rosetta InterfaceAnalyzer application (Stranges et al. 2013), these were: dSASA (buried surface area), Surface Complementarity, Packstat (Sheffler et al. 2009)

and predicted affinity. The fifth feature was the ZDOCK statistical potential and the sixth was a custom antibody-specific statistical potential (see methods Section 5.2.7). The objective ranking score was calculated by comparing the antibody-antigen interactions between the pose and the true complex. The ranking function was trained and validated in a take-one-out framework, training the ranking function on three of our lysozyme binders and testing it on the fourth (see Table 5.4).

4TTD	**Ionic**	**Hydrogen bond**	**Van-der Waals**
Top 100 poses	0.116	0.313	0.313
Remaining poses	0.057	0.139	0.113
	**	**	**
1VFB	**Ionic**	**Hydrogen bond**	**Van-der Waals**
Top 100 poses	0.000	0.260	0250
Remaining poses	0.000	0.163	0.100
		**	**
1P2C	**Ionic**	**Hydrogen bond**	**Van-der Waals**
Top 100 poses	0.162	0.419	0.199
Remaining poses	0.113	0.274	0.094
	**	**	**
3D9A	**Ionic**	**Hydrogen bond**	**Van-der Waals**
Top 100 poses	0.113	0.277	0.267
Remaining poses	0.033	0.130	0.116
	**	**	**

Table 5.4 The average fraction of interactions shared between the true complex and the designed poses. This table shows the average fraction of interactions of each type shared between the true pose (specified by its PDB id) and the set of designs. The Top 100 rows show the average fractions for the highest ranked 100 poses of our ranking algorithm while the

Remaining poses rows show the average fraction for all remaining poses. The results have been calculated in a take-one out fashion, training the algorithm on three complexes and testing on the fourth. The double star (**) highlights elements where the p-value (measured by Fisher's Exact Test) is below 10^{-10}. The interactions have been characterized using the Arpeggio software (Jubb et al. 2017). The complex that is a part of the 1VFB structure did not contain any ionic interactions. This table shows that our ranking algorithm can successfully find poses that contain similar interactions to the true antibody-antigen complexes.

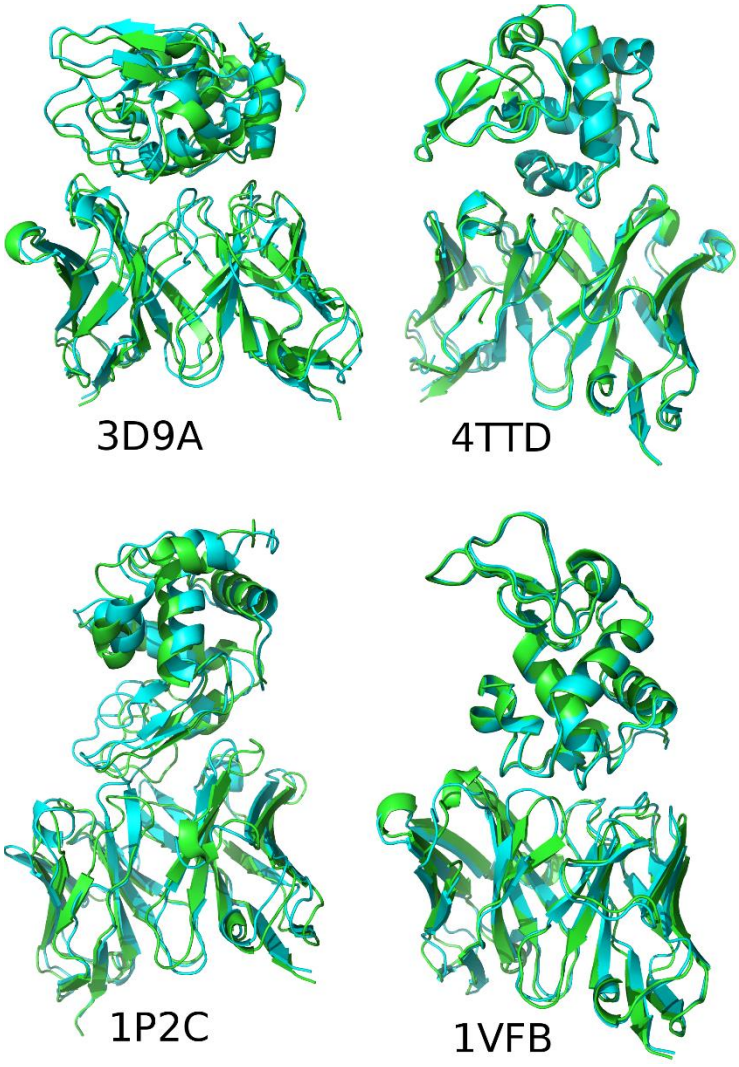

Figure 5.4 Our docking protocol recovers binding modes close to those observed in true antibody-antigen complexes. Each panel shows the structural alignment between the true pose and the most similar designed pose (as measured by TM-score) found in the set of 100 highest-

ranked poses. The true pose is shown in green and the designed pose is shown in blue. The PDB id below each panel indicates which crystal structure the true pose originates from. The figure shows that our library contains sufficient structural diversity to mimic the antibody binders with available crystal structures, and that our docking and pose selection protocol can find poses that are structurally close to the true one.

We find that the top-ranked poses preserve significantly more interactions than average, indicating that our ranking algorithm can successfully recognize docking poses that contain similar interactions to the true antibody-antigen complexes. In each of the four cases, there was at least one pose in the top 100 with a TM-score of over 0.9 to the true crystal structure (see Figure 5.4), indicating that our pose generation and selection method was able to discover binding modes similar to the structurally characterised ones, despite not being explicitly optimized to do so.

Using the pose ranking algorithm, we selected the 100 highest-ranking poses for computational affinity maturation.

5.3.6 Computational affinity maturation

The poses we generated using our model library and selected using our ranking function mirrored the structural features seen in crystal structures of antibody-antigen complexes. Nevertheless, their predicted affinity (characterized by Rosetta (Fleishman et al. 2011)) and the number of interactions were relatively low. In order to design binders capable of high affinity interactions we have created a Rosetta Scripts (Fleishman et al. 2011) protocol which simulated the hypermutation phase of antibody maturation.

To ensure that the antibody sequences we designed remain as close as possible to the available set of human sequences, we restricted the set of allowable mutations for each

CDR position within our model library (see methods Section 5.2.8) though Position Specific Substitution Tables (PSSTs). Introducing the PSSTs into our methodology ensured that our designs retained features observed in real antibody-antigen complexes and did not deviate from the human antibody sequence space (see below).

The Rosetta Scripts pipeline consisted of four steps, repeated in a Simulated Annealing (Kirkpatrick et al. 1983) trajectory for 2000 steps (see Figure 5.5), creating four repeats per each docking pose in our set of 100 highest-ranked poses, which required ~20,400 CPU hours.

5.3.7 Computationally matured lysozyme designs

The computational affinity maturation protocol was tested on the 100 top-scoring docking poses from the pool created for the 3D9A Lysozyme epitope. The protocol was followed for 2,000 Simulated Annealing steps and four trajectories were created for each docking pose, producing 400 matured designs. In parallel, we created a different set of 400 designs, without constraining the allowable substitutions by the PSST. In this second set, each selected residue could be mutated to any of the 20 amino acids. The unconstrained set was created to quantify the impact of the PSST on the designed binders.

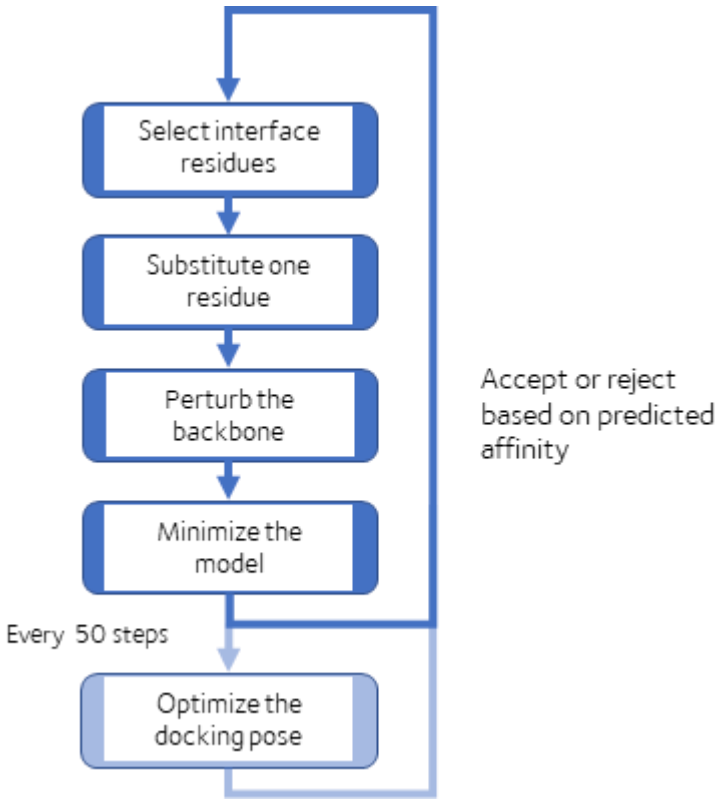

Figure 5.5 The workflow of the computational affinity maturation process. The flowchart above summarizes the steps implemented in the Rosetta Scripts protocol designed to improve the affinity of the selected poses. At each step, the protocol selects the interface residues, modifies a randomly selected interface residue, perturbs the backbone and minimizes the new model. These steps are repeated in a Simulated Annealing trajectory, where each substitution is accepted or rejected based on the change in predicted affinity. Every 50 steps the docking pose is optimized using the high-resolution phase of the Rosetta docking protocol. We ran the maturation trajectory for 2,000 steps.

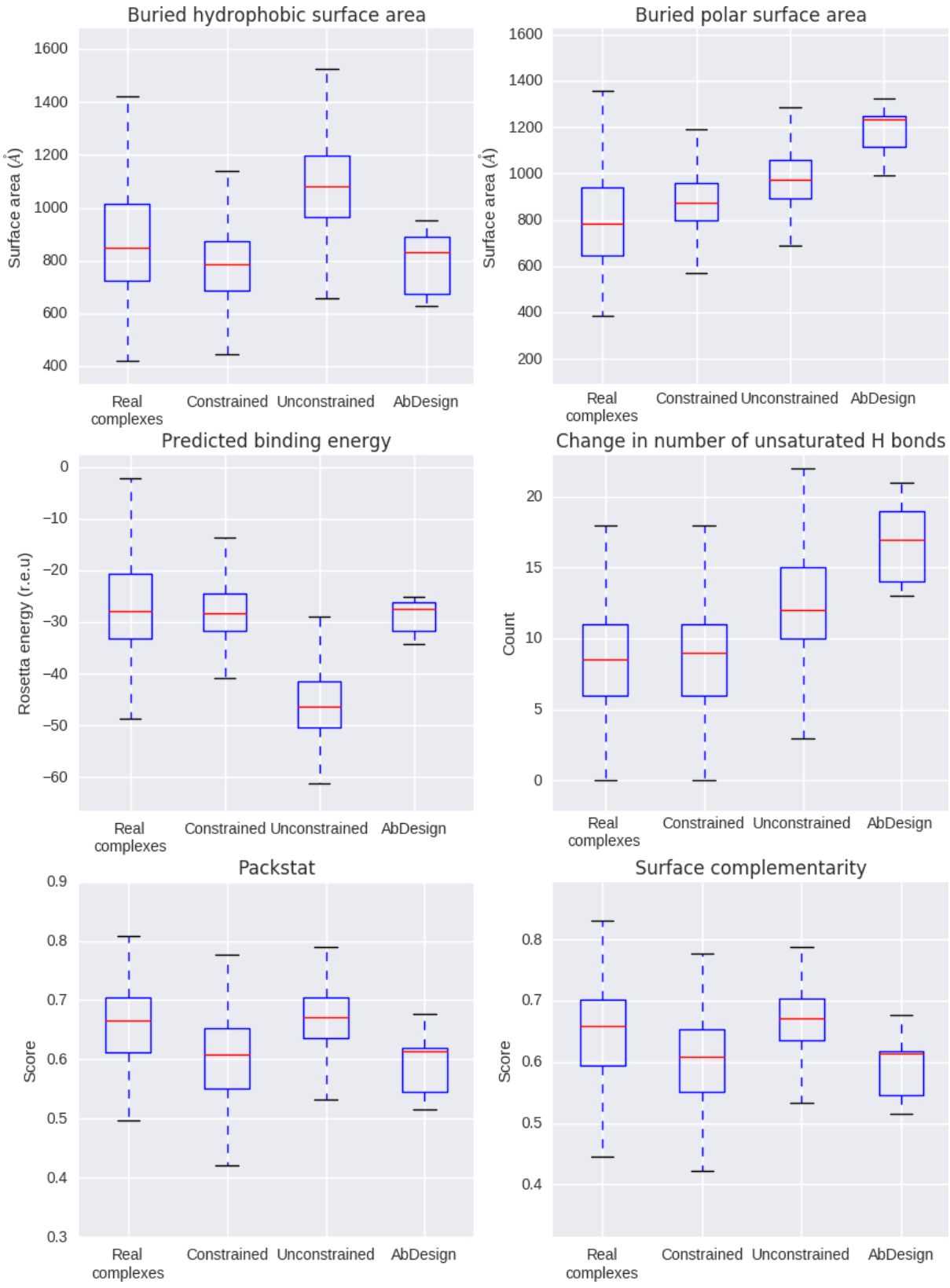

Figure 5.6 Interface quality scores comparison. The figure shows six types of Rosetta interface scores (Buried hydrophobic surface area, Buried polar surface area, Predicted binding energy,

Change in the number of unsaturated hydrogen bonds, Packstat and Surface complementarity) calculated for a set of 280 true antibody-antigen complexes, a set of 400 designs constrained by the PSST, a set of 400 unconstrained designs and a set of 12 AbDesign (Lapidoth et al. 2015) lysozyme binders. The figure shows that our PSST constrained designs which are more similar to the real antibody-antigen complexes than either the unconstrained designs or the AbDesign designs.

We compared our designs to a set of 280 antibody-antigen complexes, selected for interface diversity (see Section 5.2.7) and to a set of Lysozyme binders created by AbDesign (Lapidoth et al. 2015) software. We calculated a set of interface quality scores, using the InterfaceAnalyzer Rosetta application (Stranges et al. 2013). These are shown in Figure 5.6. The Unconstrained designs contained, on average, a large buried hydrophobic area and high numbers of unsaturated hydrogen bonds, in comparison to our constrained designs and to the real antibody-antigen complexes. The AbDesign binders contained large buried polar surface areas and, like the unconstrained designs, many unsaturated hydrogen bonds. The deviations from the properties of real antibody structures in the unconstrained and AbDesign set could mean that these designs are more likely to misfold and aggregate.

Next, we analysed the number and types of protein-protein interactions contained within our designs. To detect the interactions, we used the Arpeggio software (Jubb et al. 2017). We compared the interactions contained within both design sets (constrained and unconstrained) with the set of 280 antibody-antigen complexes (see Figure 5.7). We compared the average number of interactions for four contact types (Ionic, Hydrogen bond, Aromatic, Hydrophobic). The results show that the designs constrained by the PSST contained a similar number of interactions of each type to the real antibody-

antigen complexes. In contrast, the unconstrained designs produced large numbers of aromatic interactions and salt bridges. The propensity for a computational antibody design to introduce a large number of bulky, aromatic residues was first observed in the work of Lippow et al. (2007). The authors solved the issue by using only the electrostatic term of the energy function to estimate the binding energy. We find that using the PSST has a similar effect, by reducing the number of substitutions to aromatic residues. Separately, we found that it also reduces the number of salt bridges created by the algorithm, bringing the count of ionic interactions in line with that observed in real antibody-antigen complexes (see Figure 5.7).

After comparing the properties of the interfaces of our designs to the interfaces of real antibody-antigen complexes, we analysed the stability of our designs. First, we calculated average contribution of each CDR residue towards the total energy of the antibody (predicted by Rosetta). We compared those average contributions to a set of all available crystal structures of human antibodies (as of June 2017), the unmodified model library, the set of designs constrained by the PSST and the set of unconstrained designs (see Figure 5.8). We find that the median per-CDR-residue energy is nearly the same in the set of designs constrained by the PSST and in our unmodified models, but is higher by about 0.35 r.e.u. (Rosetta energy units) in the set of unconstrained designs. This result shows that using the PSST prevents the design protocol from reducing the predicted stability of the model library.

Next, we calculated sequence identities over CDR residues to the closest available NGS sequence and found that they are generally very close. This result shows that our design protocol finds sequences which remain near the ground state of human immune system. We also calculated the CDR identities between the NGS data and a set of 64 antibody

therapeutics with known structure. We found that the identities between the therapeutics and the NGS lie in the same range as the identities between the designs and the NGS. The results are shown in Figure 5.9.

Finally, we analysed the propensity of our designs to cause immunogenicity. In 2014 King et al. developed a computational methodology for predicting the possibility of a protein structure to cause a T-cell mediated immune response. The authors designed an immunogenicity scoring system, which they have used to redesign protein sequences of several test cases. Experimental validation showed that the designs had a significantly reduced immunogenicity, demonstrating the power of the method. We have used the scoring system of King et al to estimate the "humanness" of our designs and compare them to the set of binders created by the authors of AbDesign protocol (Lapidoth et al. 2015). We calculated the immunogenicity scores for our original library of NGS models, the set of designs constrained by PSST, the set of unconstrained designs, and the set of binders created by AbDesign (see Figure 5.10). The original library has been designed using real human antibody sequences and, therefore, can be used as a standard against which we can measure the "humanness" of the designed sequences. We find that our constrained designs have similar immunogenicity scores to the original set of NGS models. In contrast, designing the binders without the PSST constraints resulted in a significant increase in predicted immunogenicity scores. The AbDesign binders score even higher, but it is important to note that optimizing for humanness was not the goal of the AbDesign protocol and many of their designs contain framework sequences originating from non-human antibodies. These results show that our design protocol can create binders with immunogenicity scores comparable to that of native human antibody sequences.

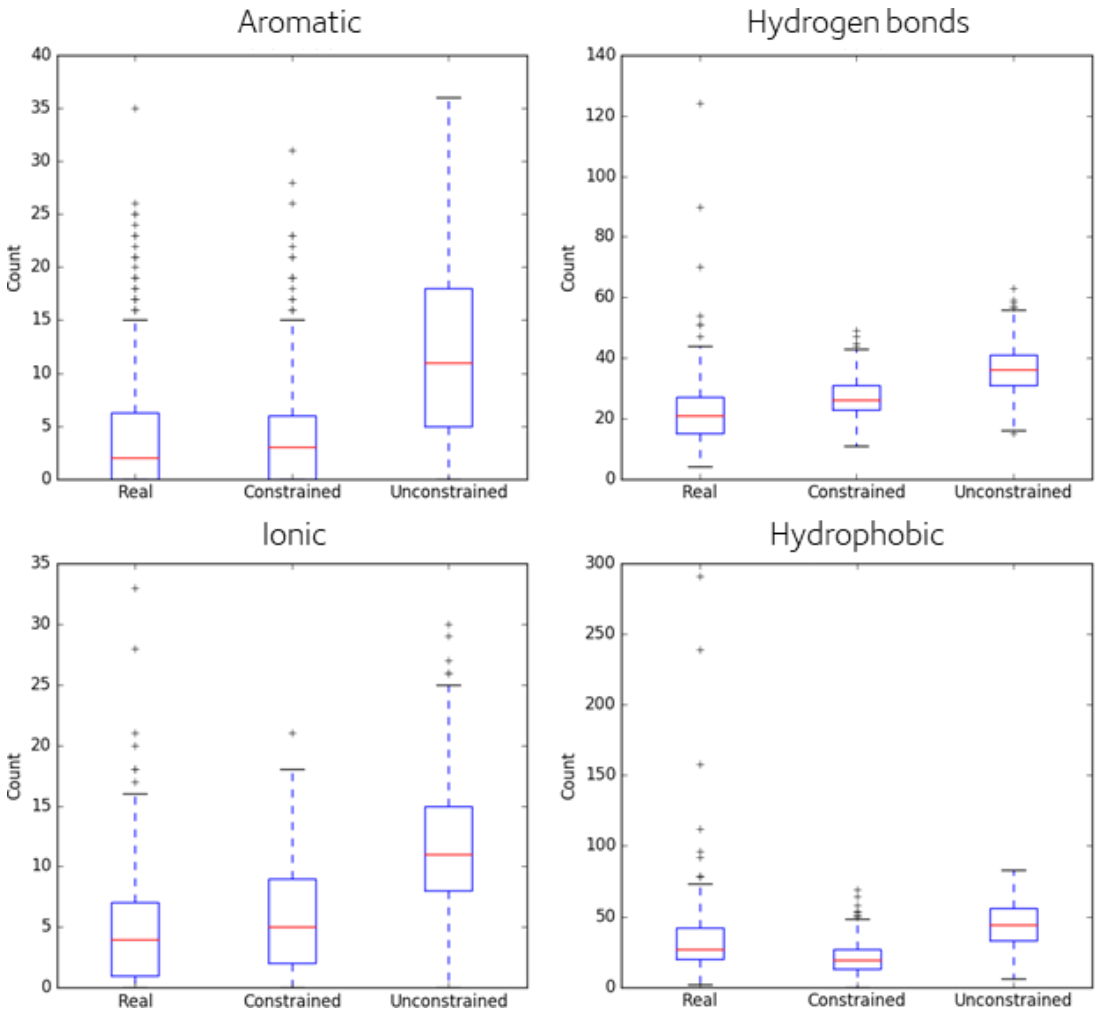

Figure 5.7 The interactions formed between the antibody and the antigen. This figure shows the average number of interactions of a given type, formed between the antibody and antigen. We compared the interaction counts between both sets of designs (constrained by PSST and unconstrained) and 280 antibody-antigen complexes from SAbDab (Dunbar et al. 2014) database. The figure shows that the constrained designs form nearly identical numbers of interactions of each type with the antigen. In contrast, the unconstrained set contains over twice as many Ionic and Aromatic interactions as the real complexes.

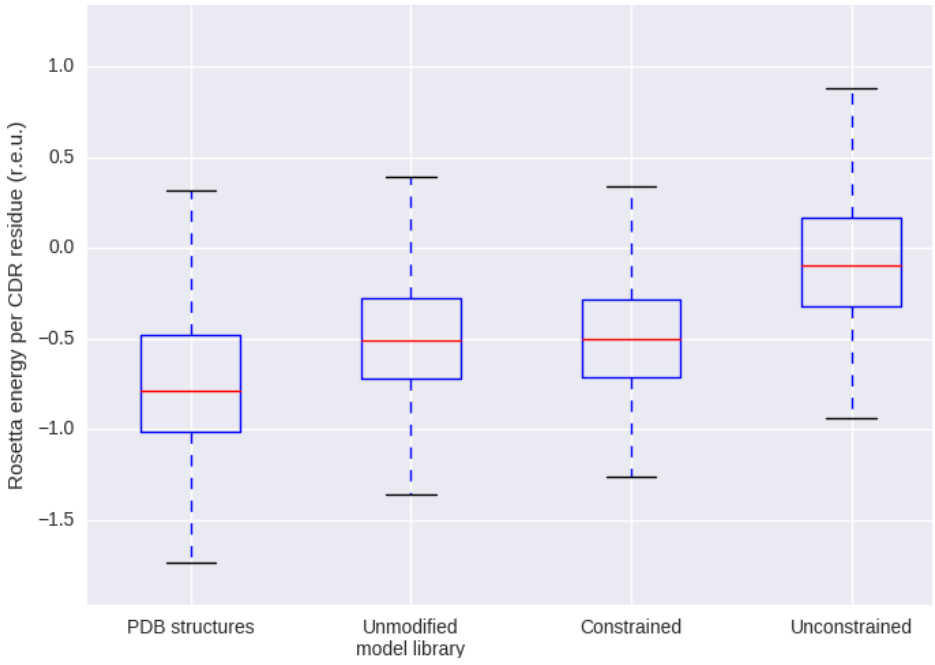

Figure 5.8 The contribution of CDR residues towards the predicted stability of the antibody. The boxplots show the average contribution of CDR residues towards the total energy of the antibody (calculated by Rosetta). The contributions were calculated for a set of 910 Human antibodies with available crystal structures, the unmodified models within our library, the set of designs constrained by PSST and the set of unconstrained designs. We find that the median energy per CDR residue is around -0.75 r.e.u. (Rosetta energy units) for the crystal structures, around -0.5 r.e.u. for the model library and for the constrained set, and around -0.15 for the unconstrained set.

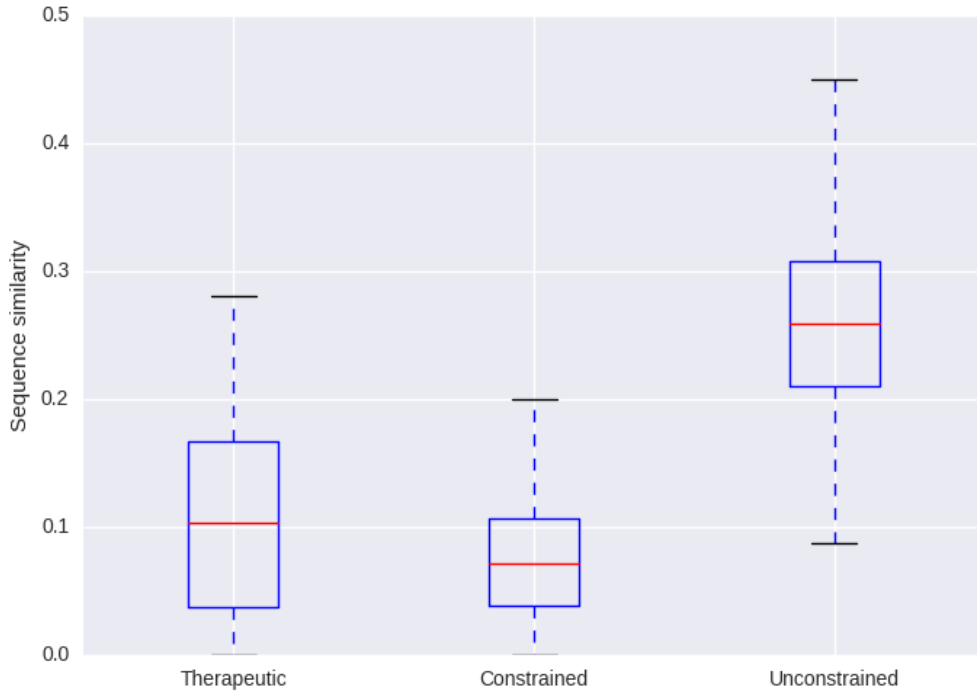

Figure 5.9 The sequence identity calculations between three sets of antibody models and the NGS data. The Figure shows the results of sequence comparisons between the NGS set and the Therapeutic antibodies, the set of PSST constrained designs and the set of unconstrained designs. The boxplots show the distribution of sequence identities, calculated over CDR residues, between the sequence of interest and the closest NGS sequence. These identities were compared for a set of 68 therapeutic antibodies with available crystal structures, the set of designs constrained by the PSST, and the set of unconstrained designs. We found that the median identity was ~10% for the therapeutic antibodies, ~7% for the constrained designs, and ~26% for the unconstrained designs. The results show that our designs remain close, in terms of sequence, to the original set of NGS sequences.

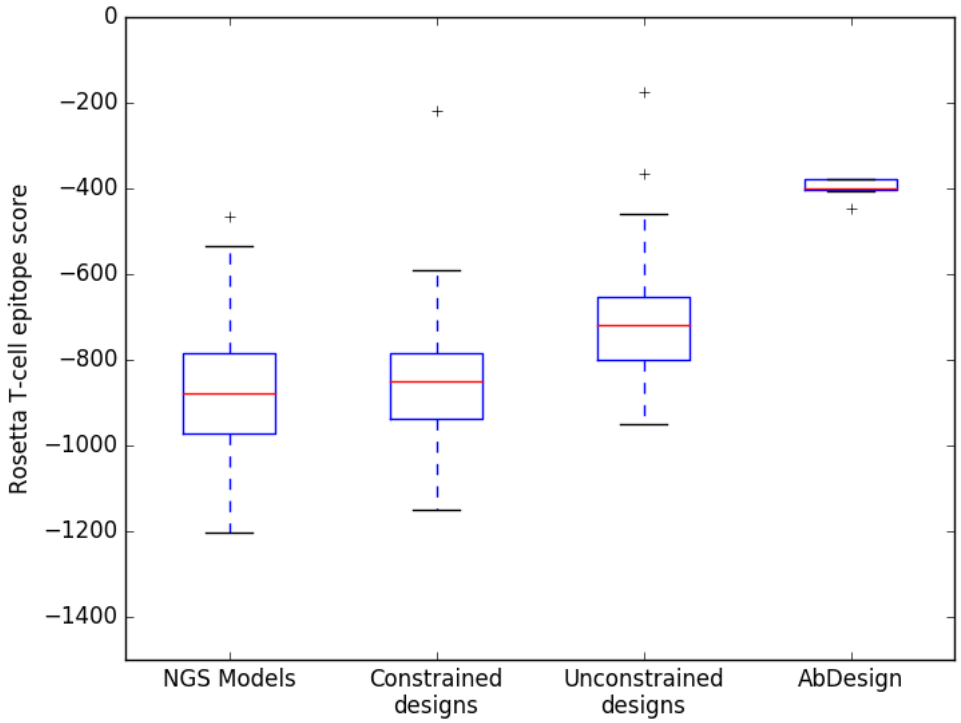

Figure 5.10 Rosetta T-cell epitope scores. The figure shows immunogenicity scores, calculated using a methodology developed by King et al(King et al. 2014), for the original library of NGS models, the set of designs constrained by the PSST, the set of unconstrained designs and the set of lysozyme binders created by the authors of AbDesign protocol (Lapidoth et al. 2015). The median score is about -900 r.e.u. (Rosetta energy units) for the NGS models, -850 r.e.u. for the constrained designs, -700 r.e.u. for the unconstrained designs and -400 r.e.u. for the AbDesign designs. The figure shows that our protocol can design binders with a low predicted propensity to cause immunogenicity.

5.4 Conclusions

In this chapter, we have described a pipeline for computational antibody design. First, we created a high-throughput structural characterization methodology, which allowed us to reduce an antibody NGS dataset containing ~15,000,000 antibody chain sequences and ~52,000,000,000,000 potential Fv sequences to the AML containing ~20,000 three-dimensional, structurally diverse Fv models. By using a dataset of human sequences, we

ensured that the library contains purely human antibodies, alleviating the immunogenicity problem. We have also observed that our models contain similar sequence and structure features to the available monoclonal antibody therapeutics, improving the chances that the library could serve as a good starting point for designing human, high affinity binders.

Next, we designed a method to computationally pan this library of models against a selected epitope of a protein target. The panning method involved creating millions of antibody-antigen docking poses and selecting 100 highest-scoring models using a machine-learning ranking algorithm. We showed that these top-scoring poses displayed many of the same interactions detected in true antibody-antigen complexes, demonstrating the power of our panning method to reproduce features observed in real antibody-antigen interfaces.

The sequences of these top-scoring poses were then redesigned using an *in-silico* affinity maturation protocol, built using the Rosetta software (Leaver-Fay, Tyka, et al. 2011). Here, we restricted the allowable mutations to the substitutions observed in the original NGS library, confining the designs inside the boundaries of human antibody sequence space. We calculated several quality metrics for our designs and showed that they fall in the range observed in real antibody-antigen complexes. We have also estimated the potential for our designs to cause immunogenicity and showed that it is no higher than in unmodified human sequences.

Overall these results show that by creating a data-driven computational antibody design pipeline, based on the results of human NGS experiments, we can develop binders that display the same features as observed in real antibody-antigen complexes. Moreover, the designs show low propensity to cause immunogenicity, improving the chances that a therapeutic designed using our method could be safely administered to humans.

The 20 designs with highest predicted affinity are currently being experimentally tested by our collaborators at UCB Pharma, with results pending.

6 CONCLUSIONS & FUTURE DIRECTIONS

6.1 Conclusions

In this section, we summarise outcomes of the research conducted for previous chapters of this thesis, and discuss the impact of these results and possible future research directions.

6.1.1 Chapter 2

In Chapter 2 we described a novel methodology for finding length-independent canonical classes of antibody CDR loops. First, we showed how we developed a dynamic programming method for comparing CDR structures of different lengths, based on Dynamic Time Warping (DTW) algorithm (Bellman et al. 1959). Using the DTW metric we showed that for ~10% of structurally characterised loops, the closest CDR is of a different length. Using this observation, we created length-independent structural clusters of CDRs. We described the clusterings for each CDR type, showing that the structures in the clusters are maintained by distinct amino acid interactions, creating strong sequence patterns, unique for each canonical class. Following that, we illustrated

how we used the discovered sequence patterns to classify novel CDR sequences into clusters. Our results show that using the length-independent approach we can classify ~20% more sequences into classes, in comparison to the standard, length-dependent approach.

Having demonstrated the practicality of our method, we investigated potential biological reasons for the observed length-independent structural similarities. The studied examples of structurally similar CDRs of different lengths indicate that they could have arisen through natural antibody diversity generating processes, such as germline sequence diversity, V(D)J recombination or somatic hypermutation. Finally, we compared our CDR clustering to recent canonical class work of North et al. (2011). We found that most of our large clusters (containing 6 or more unique sequences) map well to the work of North et al., usually having a one-to-one correspondence.

Overall, our work has demonstrated that length-independent structural similarities are an important feature of antibody space and that their prevalence could increase as new antibody structures are characterised.

The results of this Chapter have been published as a research article (Nowak et al. 2016). They also contributed to a second article (Dunbar et al. 2016).

6.1.2 Chapter 3

In the third Chapter of this thesis we considered CDR-H3 structure prediction and VH-VL orientation assignment. In the first part of the Chapter we developed a decoy ranking methodology for a loop structure prediction software Sphinx (Marks et al. 2017), created by another member of my lab, Claire Marks. First, we showed how we built and tested a machine learning Artificial Neural Network (ANN) ranking method, which

combined results produced by eight established decoy scoring functions into a consensus predictor. We found that the ANN ranking algorithm failed to significantly improve the ranking performance. We hypothesised that the bad performance of the individual methods and of the consensus function was due to a large number of steric clashes within the loop decoys. To remove those clashes we employed a Kinematic Closure (KIC) algorithm from the Rosetta package (Stein et al. 2013), which perturbed the conformation of individual decoys. We found that while the KIC protocol did not improve the RMSD of individual loop models, it did improve the performance of decoy ranking algorithms. Finally, we analysed whether crystal hydrogen bonds formed by loops in our test set have an impact on our ability to correctly predict structure of those loops. We found that loop structures forming crystal hydrogen bonds are more difficult to predict and, therefore, we recommended removing those loops from the benchmark set. The full Sphinx algorithm, including the ranking protocol, was tested on a set of 39 antibody CDR-H3 structures. In comparison to the widely-used RosettaAntibody protocol (Chaudhury et al. 2011), Sphinx produced more accurate models at a significantly reduced computational cost.

In the second part of Chapter 3, we described the development of a high-throughput VH-VL orientation assignment protocol. The algorithm assigned orientation by calculating interface sequence identity between an Fv sequence to be modelled and a set of 989 structurally-characterised orientation templates. First, we calibrated our method by finding an orientation RMSD threshold below which two orientations can be considered identical (1.5Å) and a sequence identity threshold above which the orientation can be assigned with a high accuracy (0.82). By expressing the sequence identity calculation as a sparse matrix multiplication, we significantly improved the

performance of the method, making it suitable for processing large volumes of antibody sequence data produced by Next-Generation Sequencing (NGS). We demonstrated the latter point by assigning orientation to over 5,000,000,000 Fv sequences, which required over $5x10^{12}$ sequence identity calculations.

The findings of this Chapter contributed to two publications (Marks et al. 2017; Parks et al. 2017).

6.1.3 Chapter 4

In Chapter 4 we introduced a novel method for finding sequence patterns in datasets of antibody CDR sequences. The algorithm used an autoencoder neural network (Hinton et al. 1994) which projected CDR sequences onto a low-dimensional vector of numbers and then reconstructed the sequences from this compressed representation. Parameters of the projection network were optimized by minimization of reconstruction error, calculated by comparing the original sequences with the reconstructed ones. To carry out this comparison, we used an amino acid distance matrix derived from the BLOSUM62 substitution matrix (Henikoff et al. 1992) which, as we showed, clusters together amino acids with similar chemical and physical properties. Once the low-dimensional vector representations of CDR sequences were calculated, these compressed representations were clustered using OPTICS algorithm (Ankerst et al. 1999). Sequences contained within each cluster should display distinctive sequence patterns, related to their structure.

We initially benchmarked our method on a dataset of artificial CDR sequences, with encoded sequence patterns. We then clustered these artificial sequences using the above methodology and analysed how well do the sequence patterns inside discovered clusters reflect the true amino acid relations, encoded in our dataset. The results showed

that our algorithm correctly separated the true patterns, even though a large number of sequences were classified outside of the clusters. Since we would only use one or two sequences from each cluster for experimental characterisation, the incorrect classification of sequences as lying outside main clusters would not prevent correct identification of underlying sequence patterns.

In the next section of Chapter 4, we used our algorithm to cluster CDR sequences from a large antibody NGS dataset. We clustered the CDR sequences separately for each chain type, CDR type, CDR length combination. The discovered clusters were compared to the structural groups described in Chapter 2. We found that our method correctly identified sequence patterns underlying the canonical classes discovered in Chapter 2. Nonetheless, it is important to note that in some cases structurally characterised CDR sequences from a single class were distributed over more than one sequence cluster. Analysis of the underlying data showed that this discrepancy is usually caused by a structural cluster being coded for by more than one germline subgroup. This shows that sequence patterns discovered by our algorithm can sometimes reflect underlying genetic mechanisms, instead of residue-residue interactions.

In the last section of Chapter 4, we described three sequence clusters which might be related to previously-unseen CDR canonical classes. Sequence patterns formed by CDRs falling into these clusters do not seem to be capable of sustaining residue-residue interactions observed in structurally characterised loops, suggesting they might be representative of novel structures. The CDR sequences from these three clusters are currently being structurally characterised by our collaborators at UCB Pharma, with the results pending.

6.1.4 Chapter 5

In the final Chapter of this thesis we introduced a novel pipeline for computational antibody design. First, we used an NGS antibody dataset, containing ~15,000,000 sequences of IgM chains, to construct an Antibody Model Library (AML). We filtered the NGS sequences based on our ability to structurally model all three CDRs and clustered the fully modellable sequences at 0.90 sequence identity, selecting one representative sequence from each cluster. Then, we paired the remining heavy and light sequences in an all-to-all fashion and assigned VH-VL orientation to Fv sequences using methodology described in Chapter 3. Fv sequences for which we could not assign orientation using our 0.82 interface sequence identity cut-off were discarded. We greedily clustered the remaining Fv sequences using a high-throughput structural comparison method and modelled the cluster centres using ABodyBuilder algorithm (Leem et al. 2016). The high-resolution ABodyBuilder models were structurally clustered once again, arriving at the final AML containing ~20,000 antibody models. We analysed the properties of the AML and showed that, for most of the structurally characterised therapeutics, it contained at least one highly homologous antibody model. This high structural coverage of therapeutic space improves the chances that our method will be able to discover antibodies with similar properties to known antibody drugs.

Having created the AML, we designed and benchmarked a process for panning this model library against a selected epitope of a given protein target. This panning method involved creating millions of docking poses between the antibody models and the antigen using ZDOCK software (Pierce et al. 2011) and ranking the poses using a machine learning algorithm, selecting the highest ranking 100 poses for further analysis. We showed that the ranking function we designed can correctly identify docking poses containing interactions similar to those observed in real antibody-antigen complexes, increasing the

likelihood that antibodies designed using our method will recapitulate binding modes observed in true complexes.

Finally, we computationally matured 100 poses designed for a known Lysozyme epitope. To carry out the maturation, we built a RosettaScripts protocol which introduced random mutations to the paratope of the antibody, that were accepted or rejected in a Simulated Annealing trajectory. The allowed mutations were constrained to amino acids observed in the original NGS data, using a Position Specific Substitution Table (PSST). We showed that our matured designs have similar features to real antibody-antigen complexes and have low predicted propensity to cause immunogenic response. These final results show that by using our PSST we can constrain the designs within the space of observed human antibody binders, making it more probable that therapeutics designed using our pipeline would strongly bind to the designated epitope and not cause immunogenic response once administered.

The matured designs with highest predicted affinity are currently being experimentally characterised by our collaborators at UCB Pharma, with the results pending.

6.2 Future work

In this section, we discuss potential future avenues for research conducted for this thesis.

6.2.1 CDR canonical classes

As mentioned before, there has been a plethora of studies into canonical classes formed by antibody CDRs. Nevertheless, most of these studies concerned themselves with the contemporary "snapshot" of the available structural data and quickly became obsolete as new antibody structures were solved. Therefore, it would be beneficial to create an

online resource, possibly as part of the SAbDab database, which would constantly update the CDR clustering as new structures arrive. This way, researchers could always access canonical class assignments for most of the structurally characterised CDRs.

The speed of canonical class assignment makes the concept uniquely suited towards structural modelling of large volumes of antibody NGS data. The investigation carried out in Chapter 2 could therefore be extended by researching different class assignment methods, towards optimal speed and accuracy. The accuracy of class assignment could also be improved by conditioning the assignment algorithms not only on structurally characterised sequences, but also on unlabelled NGS data, using the sequence clusters found in Chapter 4.

6.2.2 Structure prediction of CDR-H3 and VH-VL orientation

We showed that the Sphinx software is able to produce accurate structural models of CDR-H3 loops. Nevertheless, at times the method failed to correctly predict the structure. More research is required into the conditions which prevent accurate modelling of the CDR-H3 loop. Currently, very few structure prediction methods indicate how much confidence should be put into the produced models, despite the importance of this information. As shown in Chapter 5, when conducting a large-scale structural investigation of a sequence dataset, often a structural prediction is not required for every data point, as it is more beneficial to have a smaller set of high-confidence models.

The VH-VL orientation assignment method we created allowed us to process billions of VH-VL sequences in tractable time. Nevertheless, we found that the coverage of the sequence space is relatively low, in that we could make a confident prediction in only ~0.1% of cases. As an extension of the work, more research could be conducted into

increasing this coverage without losing the accuracy and speed of prediction. For example, different interface residues could be assigned different weights in the identity calculations, depending on their predicted importance for maintaining the VH-VL orientation. Another avenue of research could involve conditioning the orientation predictors using NGS antibody datasets. The work of DeKosky et al. (2014) and others into sequencing paired VH-VL repertoire led to development of methods for experimental characterisation of full Fv sequences. Datasets of such paired heavy-light sequences could be analysed using contact prediction algorithms (Morcos et al. 2011) to determine residue pairs crucial for maintaining the orientation.

6.2.3 Determination of CDR sequence patterns

The method we have developed in Chapter 4 allowed us to discover clusters of CDR sequences containing novel patterns, potentially indicative of unseen canonical classes. If the experimental validation confirms our hypothesis, we could submit more sequence clusters for structural characterisation, extending our knowledge of antibody conformational space.

In our autoencoder architecture we encoded the amino acids sequences using one-hot coding. This was done to encourage sparsity in the projection layer. Nevertheless, it would be worthwhile to explore other representation methods, such as transforming the sequences into vectors describing physicochemical and biochemical features of constituent amino acids. The features could be taken from an external database such as the Amino Acid Index Database (AAIndex) (Kawashima et al. 1999).

We implemented the reconstruction cost in our autoencoder neural network using a BLOSUM62 substitution matrix. A possible avenue of future research could be to investigate using other substitution models in the architecture. Such substitution model

could be antibody specific, such as the one developed by Mirsky et al. 2015. Given enough data, one could derive their own substitution models which could be germline-specific, CDR-specific or even position-specific. The algorithm could also be extended to other protein families, using appropriate family-specific substitution models (e.g. Sau et al. 2011).

The pattern discovery algorithm requires a large amount of sequence data to be able to discern the residue relationships from background noise. In the work conducted in Chapter 4 we used only one NGS dataset of human IgM sequences. By collating data obtained from other publicly and privately available resources, we could put more confidence in our results and analyse sequence groups which are currently sparsely populated. With more sequence data, we could also extend our analysis into CDR-H3 sequences and investigate whether the loop could be capable of forming canonical classes, which remains an unanswered question.

Antibody sequences from non-human immune systems often belong to completely different structural space than the human ones. Using our method, these differences could be quantified more rigorously, providing an insight into the uniqueness of human immune system. Such investigation could have an impact on humanization efforts, by giving a measure of equivalence between CDR sequences of different species.

6.2.4 Data-driven antibody design

In Chapter 5 we showed that the antibody structures designed using our method had similar features to real antibody therapeutics. The designed binders are currently being investigated by our collaborators at UCB Pharma. If the designs are shown to bind their target at high affinity and specificity, it would constitute a very strong validation of our design strategy. We could then move onto more therapeutically relevant targets,

comparing the properties of our designs to existing mAbs. If the designs fail to strongly bind their target, the experiments could still provide valuable insights on why that happens. The algorithm could then be modified to reflect those observations and new rounds of designs could be tested.

In Chapter 5, we focused on designing antibodies with human properties, to avoid immunogenic response against the binder. Nevertheless, there are non-human antibodies with non-standard structures which have recently been investigated by therapeutic researchers, such as single-chain camelid nanobodies or bovine antibodies. If relevant NGS data could be obtained, our algorithm could be readily modified to design binders with the aforementioned structural properties. By combining the human data with camelid data, the pipeline could also be steered towards designing soluble heavy chains with reduced propensity to cause immunogenic response.

The steps involved in our design pipeline required large amounts of computational resources. By streamlining the computation, the process could be accelerated, making it easier to create a large number of candidate binders, which would then be selected for experimental validation.

7 REFERENCES

Aalberse, R. C. and Schuurman, J. 2002. "IgG4 Breaking the Rules." *Immunology* 105(1):9–19.

Abadi, M., Agarwal, A., Paul, B., Eugene, B., Zhifeng, C., Craig, C., Greg, C. S., Davis, A., Dean, J., Devin, M., et al. 2015. "TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems."

Abhinandan, K. R. and Martin, A. C. R. 2008. "Analysis and Improvements to Kabat and Structurally Correct Numbering of Antibody Variable Domains." *Molecular Immunology* 45(14):3832–39.

Abhinandan, K. R. and Martin, A. C. R. 2010. "Analysis and Prediction of VH/VL Packing in Antibodies." *Protein Engineering Design and Selection* 23(9):689–97.

Adolf-Bryfogle, J., Xu, Q., North, B., Lehmann, A., and Dunbrack Jr., R. L. 2015. "PyIgClassify: A Database of Antibody CDR Structural Classifications." *Nucleic Acids Research* 43(Database-Issue):432–38.

Al-Lazikani, B., Lesk, A. M., and Chothia, C. 1997. "Standard Conformations for the Canonical Structures of Immunoglobulins." *Journal of Molecular Biology* 273(4):927–48.

Almagro, J. C. and Fransson, J. 2008. "Humanization of Antibodies." *Frontiers in Bioscience : A Journal and Virtual Library* 13:1619–33.

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. "Gapped BLAST and PSI-BLAST: A New Generation of Protein Database Search Programs." *Nucleic Acids Research* 25(17):3389–3402.

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

Babor, M., Mandell, D. J., and Kortemme, T. 2011. "Assessment of Flexible Backbone Protein Design Methods for Sequence Library Prediction in the Therapeutic Antibody Herceptin-HER2 Interface." *Protein Science : A Publication of the Protein Society* 20(6):1082–89.

Barclay, A. N. 1999. "Ig-like Domains: Evolution from Simple Interaction Molecules to

Sophisticated Antigen Recognition." *Proceedings of the National Academy of Sciences of the United States of America* 96(26):14672–74.

Barre, S., Greenberg, A. S., Flajnik, M. F., and Chothia, C. 1994. "Structural Conservation of Hypervariable Regions in Immunoglobulins Evolution." *Nature Structural Biology* 1(12):915–20.

Bellman, R. and Kalaba, R. 1959. "On Adaptive Control Processes." *IRE Transactions on Automatic Control* 4(2):1–9.

Berg, J. M., Tymoczko, J. L., and Stryer, L. 2012. *Biochemistry: International Edition*. Basingstoke: W.H. Freeman.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. 2000. "The Protein Data Bank." *Nucleic Acids Res* 28(1):235–42.

Bork, P. and Koonin, E. V. 1996. "Protein Sequence Motifs." *Current Opinion in Structural Biology* 6(3):366–76.

Brezinschek, H. P., Foster, S. J., Dörner, T., Brezinschek, R. I., and Lipsky, P. E. 1998. "Pairing of Variable Heavy and Variable Kappa Chains in Individual Naive and Memory B Cells." *Journal of Immunology* 160(10):4762–67.

Briney, B. S., Willis, J. R., and Crowe, J. E. 2012. "Location and Length Distribution of

Somatic Hypermutation-Associated DNA Insertions and Deletions Reveals Regions of Antibody Structural Plasticity." *Genes and Immunity* 13(7):523–29.

Van de Broek, B., Devoogdt, N., D'Hollander, A., Gijs, H.-L., Jans, K., Lagae, L., Muyldermans, S., Maes, G., and Borghs, G. 2011. "Specific Cell Targeting with Nanobody Conjugated Branched Gold Nanoparticles for Photothermal Therapy." *ACS Nano* 5(6):4319–28.

Bronstein, A. M., Bronstein, M. M., and Kimmel, R. 2006. "Generalized Multidimensional Scaling: A Framework for Isometry-Invariant Partial Surface Matching." *Proceedings of the National Academy of Sciences* 103(5):1168–72.

Brunger, A. T., DeLabarre, B., Davies, J. M., and Weis, W. I. 2009. "X-Ray Structure Determination at Low Resolution." *Acta Crystallographica Section D: Biological Crystallography* 65(2):128–33.

Bujotzek, A., Dunbar, J., Lipsmeier, F., Schäfer, W., Antes, I., Deane, C. M., and Georges, G. 2015. "Prediction of VH-VL Domain Orientation for Antibody Variable Domain Modeling." *Proteins* 83(4):681–95.

Burges, C. J. C. 2010. *From RankNet to LambdaRank to LambdaMART: An Overview*.

Burkovitz, A. and Ofran, Y. 2016. "Understanding Differences between Synthetic and Natural Antibodies Can Help Improve Antibody Engineering." *mAbs* 8(2):278–87.

Canutescu, A. A. and Dunbrack, R. L. 2003. "Cyclic Coordinate Descent: A Robotics Algorithm for Protein Loop Closure." *Protein Science : A Publication of the Protein Society* 12(5):963–72.

Carugo, O. and Pongor, S. 2001. "A Normalized Root-Mean-Square Distance for Comparing Protein Three-Dimensional Structures." *Protein Science : A Publication of the Protein Society* 10(7):1470–73.

Chaffey, N., Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. 2003. "Molecular Biology of the Cell." *Annals of Botany* 91(3):401.

Chailyan, A., Marcatili, P., Cirillo, D., and Tramontano, A. 2011. "Structural Repertoire of Immunoglobulin λ Light Chains." *Proteins: Structure, Function, and Bioinformatics* 79(5):1513–24.

Chailyan, A., Tramontano, A., and Marcatili, P. 2012. "A Database of Immunoglobulins with Integrated Tools: DIGIT." *Nucleic Acids Research* 40(Database issue):D1230-4.

Chames, P. and Baty, D. 2009. "Bispecific Antibodies for Cancer Therapy: The Light at the End of the Tunnel?" *mAbs* 1(6):539–47.

Chapelle, O. 2011. "Yahoo! Learning to Rank Challenge Overview." *JMLR: Workshop and Conference Proceedings* 14:1–24.

Chaudhury, S., Berrondo, M., Weitzner, B. D., Muthu, P., Bergman, H., and Gray, J. J. 2011. "Benchmarking and Analysis of Protein Docking Performance in Rosetta v3.2" edited by V. N. Uversky. *PLoS ONE* 6(8):e22477.

Chen, W., Liu, T., Lan, Y., Ma, Z., and Li, H. 2009. "Ranking Measures and Loss Functions in Learning to Rank." *Nips* 1–9.

Choi, Y. and Deane, C. M. 2010. "FREAD Revisited: Accurate Loop Structure Prediction Using a Database Search Algorithm." *Proteins* 78(6):1431–40.

Chothia, C. and Lesk, A. M. 1986. "The Relation between the Divergence of Sequence and Structure in Proteins." *The EMBO Journal* 5(4):823–26.

Chothia, C. and Lesk, A. M. 1987. "Canonical Structures for the Hypervariable Regions of Immunoglobulins." *Journal of Molecular Biology* 196(4):901–17.

Chothia, C., Lesk, A. M., Gherardi, E., Tomlinson, I. M., Walter, G., Marks, J. D., Llewelyn, M. B., and Winter, G. 1992. "Structural Repertoire of the Human VH Segments." *Journal of Molecular Biology* 227(3):799–817.

Chothia, C., Lesk, A. M., Tramontano, A., Levitt, M., Smith-Gill, S. J., Air, G., Sheriff, S., Padlan, E. A., Davies, D., Tulip, W. R., et al. 1989. "Conformations of Immunoglobulin Hypervariable Regions." *Nature* 342(6252):877–83.

Clark, L. A., Boriack-Sjodin, P. A., Day, E., Eldredge, J., Fitch, C., Jarpe, M., Miller, S., Li, Y.,

Simon, K., and van Vlijmen, H. W. T. 2009. "An Antibody Loop Replacement Design Feasibility Study and a Loop-Swapped Dimer Structure." *Protein Engineering, Design & Selection : PEDS* 22(2):93–101.

Clark, L. A., Boriack-Sjodin, P. A., Eldredge, J., Fitch, C., Friedman, B., Hanf, K. J. M., Jarpe, M., Liparoto, S. F., Li, Y., Lugovskoy, A., et al. 2006. "Affinity Enhancement of an in Vivo Matured Therapeutic Antibody Using Structure-Based Computational Design." *Protein Science : A Publication of the Protein Society* 15(5):949–60.

Colman, P. M., Laver, W. G., Varghese, J. N., Baker, A. T., Tulloch, P. A., Air, G. M., and Webster, R. G. 1987. "Three-Dimensional Structure of a Complex of Antibody with Influenza Virus Neuraminidase." *Nature* 326(6111):358–63.

Conway, P., Tyka, M. D., DiMaio, F., Konerding, D. E., and Baker, D. 2014. "Relaxation of Backbone Bond Geometry Improves Protein Energy Landscape Modeling." *Protein Science* 23(1):47–55.

Crooks, G. E., Hon, G., Chandonia, J.-M., and Brenner, S. E. 2004. "WebLogo: A Sequence Logo Generator." *Genome Research* 14(6):1188–90.

DeKosky, B. J., Kojima, T., Rodin, A., Charab, W., Ippolito, G. C., Ellington, A. D., and Georgiou, G. 2014. "In-Depth Determination and Analysis of the Human Paired Heavy- and Light-Chain Antibody Repertoire." *Nature Medicine* 21(1):86–91.

Denning, P. J. and Lewis, T. G. 2016. "Exponential Laws of Computing Growth." *Communications of the ACM* 60(1):54–65.

Dong, G. Q., Fan, H., Schneidman-Duhovny, D., Webb, B., and Sali, A. 2013. "Optimized Atomic Statistical Potentials: Assessment of Protein Interfaces and Loops." *Bioinformatics (Oxford, England)* 29(24):3158–66.

Dunbar, J. and Deane, C. M. 2015. "ANARCI: Antigen Receptor Numbering and Receptor Classification." *Bioinformatics* 32(2):298–300.

Dunbar, J., Fuchs, A., Shi, J., and Deane, C. M. 2013. "ABangle: Characterising the VH-VL Orientation in Antibodies." *Protein Engineering, Design & Selection : PEDS* 26(10):611–20.

Dunbar, J., Krawczyk, K., Leem, J., Baker, T., Fuchs, A., Georges, G., Shi, J., and Deane, C. M. 2014. "SAbDab: The Structural Antibody Database." *Nucleic Acids Research* 42(D1):D1140-6.

Dunbar, J., Krawczyk, K., Leem, J., Marks, C., Nowak, J., Regep, C., Georges, G., Kelm, S., Popovic, B., and Deane, C. M. 2016. "SAbPred: A Structure-Based Antibody Prediction Server." *Nucleic Acids Research* 44(W1):W474–78.

Eddy, S. R. 1998. "Profile Hidden Markov Models." *Bioinformatics* 14(9):755–63.

Edmundson, A. B., Harris, D. L., Fan, Z. C., Guddat, L. W., Schley, B. T., Hanson, B. L.,

Tribbick, G., and Geysen, H. M. 1993. "Principles and Pitfalls in Designing Site-Directed Peptide Ligands." *Proteins: Structure, Function, and Bioinformatics* 16(3):246–67.

Elvin, J. G., Couston, R. G., and van der Walle, C. F. 2013. "Therapeutic Antibodies: Market Considerations, Disease Targets and Bioprocessing." *International Journal of Pharmaceutics* 440(1):83–98.

Ester, M., Kriegel, H.-P., Sander, J., and Xu, X. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." Pp. 226–31 in *Proc. of 2nd International Conference on Knowledge Discovery and Data Mining*.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., Pieper, U., and Sali, A. 2006. "Comparative Protein Structure Modeling Using Modeller." *Current Protocols in Bioinformatics* Chapter 5:Unit 5.6.

Fasnacht, M., Butenhof, K., Goupil-Lamy, A., Hernandez-Guzman, F., Huang, H., and Yan, L. 2014. "Automated Antibody Structure Prediction Using Accelrys Tools: Results and Best Practices." *Proteins: Structure, Function, and Bioinformatics* 82(8):1583–98.

Finkel, J. R., Kleeman, A., and Manning, C. D. 2008. "Efficient, Feature-Based, Conditional Random Field Parsing." Pp. 959–67 in *Proceedings of ACL-08: HLT*. Columbus, Ohio: Association for Computational Linguistics.

Fleishman, S. J., Leaver-Fay, A., Corn, J. E., Strauch, E. M., Khare, S. D., Koga, N., Ashworth, J., Murphy, P., Richter, F., Lemmon, G., et al. 2011. "Rosettascripts: A Scripting Language Interface to the Rosetta Macromolecular Modeling Suite." *PLoS ONE* 6(6):e20161.

Friedensohn, S., Khan, T. A., and Reddy, S. T. 2017. "Advanced Methodologies in High-Throughput Sequencing of Immune Repertoires." *Trends in Biotechnology* 35(3):203–14.

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. 2012. "CD-HIT: Accelerated for Clustering the next-Generation Sequencing Data." *Bioinformatics (Oxford, England)* 28(23):3150–52.

Furukawa, K., Shirai, H., Azuma, T., and Nakamura, H. 2001. "A Role of the Third Complementarity-Determining Region in the Affinity Maturation of an Antibody." *Journal of Biological Chemistry* 276(29):22–28.

Galitsky, B. A., Gelfand, I. M., and Kister, A. E. 1998. "Predicting Amino Acid Sequences of the Antibody Human VH Chains from Its First Several Residues." *Proceedings of the National Academy of Sciences of the United States of America* 95(9):5193–98.

Galson, J. D., Trück, J., Clutterbuck, E. A., Fowler, A., Cerundolo, V., Pollard, A. J., Lunter, G., and Kelly, D. F. 2016. "B-Cell Repertoire Dynamics after Sequential Hepatitis B

Vaccination and Evidence for Cross-Reactive B-Cell Activation." *Genome Medicine* 8(1):68.

Giudicelli, V., Duroux, P., Ginestoux, C., Folch, G., Jabado-Michaloud, J., Chaume, D., and Lefranc, M.-P. 2006. "IMGT/LIGM-DB, the IMGT Comprehensive Database of Immunoglobulin and T Cell Receptor Nucleotide Sequences." *Nucleic Acids Research* 34(Database issue):D781-4.

Glanville, J., Zhai, W., Berka, J., Telman, D., Huerta, G., Mehta, G. R., Ni, I., Mei, L., Sundar, P. D., Day, G. M. R., et al. 2009. "Precise Determination of the Diversity of a Combinatorial Antibody Library Gives Insight into the Human Immunoglobulin Repertoire." *Proceedings of the National Academy of Sciences* 106(48):20216–21.

Goldman, E. R., Liu, J. L., Zabetakis, D., and Anderson, G. P. 2017. "Enhancing Stability of Camelid and Shark Single Domain Antibodies: An Overview." *Frontiers in Immunology* 8:865.

Gray, J. J., Moughon, S., Wang, C., Schueler-Furman, O., Kuhlman, B., Rohl, C. A., and Baker, D. 2003. "Protein–Protein Docking with Simultaneous Optimization of Rigid-Body Displacement and Side-Chain Conformations." *Journal of Molecular Biology* 331(1):281–99.

Green, L. L., Hardy, M. C., Maynard-Currie, C. E., Tsuda, H., Louie, D. M., Mendez, M. J.,

Abderrahim, H., Noguchi, M., Smith, D. H., Zeng, Y., et al. 1994. "Antigen–specific Human Monoclonal Antibodies from Mice Engineered with Human Ig Heavy and Light Chain YACs." *Nature Genetics* 7(1):13–21.

Grey, J. L. and Thompson, D. H. 2010. "Challenges and Opportunities for New Protein Crystallization Strategies in Structure-Based Drug Design." *Expert Opinion on Drug Discovery* 5(11):1039–45.

Hahnloser, R. H. R., Sarpeshkar, R., Mahowald, M. A., Douglas, R. J., and Seung, H. S. 2000. "Digital Selection and Analogue Amplification Coexist in a Cortex-InspiredSilicon Circuit." *Nature* 405(6789):947–51.

Harris, D. M. and Harris, S. L. 2013. *Digital Design and Computer Architecture*. Morgan Kaufmann.

Harwood, N. E. and Batista, F. D. 2010. "Early Events in B Cell Activation." *Annual Review of Immunology* 28(1):185–210.

Henikoff, J. G. and Henikoff, S. 1996. "Blocks Database and Its Applications." *Methods Enzymol* 266:88–105.

Henikoff, S. and Henikoff, J. G. 1992. "Amino Acid Substitution Matrices from Protein Blocks." *Proceedings of the National Academy of Sciences* 89(22):10915–19.

Hildebrand, P. W., Goede, A., Bauer, R. A., Gruening, B., Ismer, J., Michalsky, E., and

Preissner, R. 2009. "SuperLooper--a Prediction Server for the Modeling of Loops in Globular and Membrane Proteins." *Nucleic Acids Research* 37(Web Server issue):W571-4.

Hinton, G. E. and Richard, Z. S. 1994. "Autoencoders, Minimum Description Length and Helmholtz Free Energy." Pp. 3–10 in *Advances in Neural Information Processing Systems*, vol. 3.

Holtby, D., Li, S. C., and Li, M. 2013. "LoopWeaver: Loop Modeling by the Weighted Scaling of Verified Proteins." *Journal of Computational Biology : A Journal of Computational Molecular Cell Biology* 20(3):212–23.

Honegger, A. and Plückthun, A. 2001. "Yet Another Numbering Scheme for Immunoglobulin Variable Domains: An Automatic Modeling and Analysis Tool." *Journal of Molecular Biology* 309(3):657–70.

Hu, S., Liang, S., Guo, H., Zhang, D., Li, H., Wang, X., Yang, W., Qian, W., Hou, S., Wang, H., et al. 2013. "Comparison of the Inhibition Mechanisms of Adalimumab and Infliximab in Treating Tumor Necrosis Factor α-Associated Diseases from a Molecular View." *Journal of Biological Chemistry* 288(38):27059–67.

Hubbard, S. J. and Thornton, J. M. 1993. "Naccess Computer Program." *Department of Biochemistry and Molecular Biology, University College London* 2(1).

Hudis, C. A. 2007. "Trastuzumab-Mechanism of Action and Use in Clinical Practice." *The New England Journal of Medicine* 357(1):39–51.

Hussack, G., Arbabi-Ghahroudi, M., van Faassen, H., Songer, J. G., Ng, K. K.-S., MacKenzie, R., and Tanha, J. 2011. "Neutralization of *Clostridium Difficile* Toxin A with Single-Domain Antibodies Targeting the Cell Receptor Binding Domain." *Journal of Biological Chemistry* 286(11):8961–76.

Huston, J. S., Levinson, D., Mudgett-Hunter, M., Tai, M. S., Novotný, J., Margolies, M. N., Ridge, R. J., Bruccoleri, R. E., Haber, E., Crea, R., et al. 1988. "Protein Engineering of Antibody Binding Sites: Recovery of Specific Activity in an Anti-Digoxin Single-Chain Fv Analogue Produced in Escherichia Coli." *Proceedings of the National Academy of Sciences of the United States of America* 85(16):5879–83.

Jackson, K. J. L., Kidd, M. J., Wang, Y., and Collins, A. M. 2013. "The Shape of the Lymphocyte Receptor Repertoire: Lessons from the B Cell Receptor." *Frontiers in Immunology* 4:263.

Jayaram, N., Bhowmick, P., and Martin, A. C. R. 2012. "Germline VH/VL Pairing in Antibodies." *Protein Engineering, Design & Selection : PEDS* 25(10):523–29.

Jeffreys, H. 1946. "An Invariant Form for the Prior Probability in Estimation Problems." *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 186(1007):453–61.

Joachims, T. 2002. "Optimizing Search Engines Using Clickthrough Data." P. 133 in *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*. New York, New York, USA: ACM Press.

Johnson, G. and Wu, T. T. 2001. "Kabat Database and Its Applications: Future Directions." *Nucleic Acids Research* 29(1):205–6.

Jubb, H. C., Higueruelo, A. P., Ochoa-Montaño, B., Pitt, W. R., Ascher, D. B., and Blundell, T. L. 2017. "Arpeggio: A Web Server for Calculating and Visualising Interatomic Interactions in Protein Structures." *Journal of Molecular Biology* 429(3):365–71.

Kabat, E. A. and National Institutes of Health. 1983. *Sequences of Proteins of Immunological Interest*. Bethesda: National Institutes of Health.

Kabsch, W. 1976. "A Solution for the Best Rotation to Relate Two Sets of Vectors." *Acta Crystallographica Section A: Crystal Physics, Diffraction, Theoretical and General Crystallography* 32(5):922–23.

Kabsch, W. and Sander, C. 1983. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." *Biopolymers* 22(12):2577–2637.

Karpusas, M., Ferrant, J., Weinreb, P. H., Carmillo, A., Taylor, F. R., and Garber, E. A. 2003. "Crystal Structure of the alpha1beta1 Integrin I Domain in Complex with an

Antibody Fab Fragment." *Journal of Molecular Biology* 327(5):1031–41.

Kashmiri, S. V. S., De Pascalis, R., Gonzales, N. R., and Schlom, J. 2005. "SDR Grafting - A New Approach to Antibody Humanization." *Methods* 36(1):25–34.

Katchalski-Katzir, E., Shariv, I., Eisenstein, M., Friesem, A. A., Aflalo, C., and Vakser, I. A. 1992. "Molecular Surface Recognition: Determination of Geometric Fit between Proteins and Their Ligands by Correlation Techniques." *Proceedings of the National Academy of Sciences of the United States of America* 89(6):2195–99.

Kawashima, S., Ogata, H., and Kanehisa, M. 1999. "AAindex: Amino Acid Index Database." *Nucleic Acids Research* 27(1):368–69.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. 1958. "A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis." *Nature* 181(4610):662–66.

Kennett, R. G. 1979. "Monoclonal Antibodies. Hybrid Myelomas--a Revolution in Serology and Immunogenetics." *American Journal of Human Genetics* 31(5):539–47.

King, C., Garza, E. N., Mazor, R., Linehan, J. L., Pastan, I., Pepper, M., and Baker, D. 2014. "Removing T-Cell Epitopes with Computational Protein Design." *Proceedings of the National Academy of Sciences of the United States of America* 111(23):8577–82.

Kingma, D. P. and Ba, J. 2014. "Adam: A Method for Stochastic Optimization." *CoRR* abs/1412.6.

Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. 1983. "Optimization by Simulated Annealing." *Science* 220(4598).

Kiyoshi, M., Caaveiro, J. M. M., Miura, E., Nagatoishi, S., Nakakido, M., Soga, S., Shirai, H., Kawabata, S., and Tsumoto, K. 2014. "Affinity Improvement of a Therapeutic Antibody by Structure-Based Computational Design: Generation of Electrostatic Interactions in the Transition State Stabilizes the Antibody-Antigen Complex." *PloS One* 9(1):e87099.

Kleywegt, G. J. and Jones, T. A. 1997. "Model Building and Refinement Practice." *Methods in Enzymology* 277:208–30.

Köhler, G. and Milstein, C. 1975. "Continuous Cultures of Fused Cells Secreting Antibody of Predefined Specificity." *Nature* 256(5517):495–97.

Kotz, S., Johnson, N. L., Balakrishnan, N., and Johnson, N. L. 2000. *Continuous Multivariate Distributions. Vol. 1, Models and Applications.* Wiley.

Krawczyk, K., Baker, T., Shi, J., and Deane, C. M. 2013. "Antibody I-Patch Prediction of the Antibody Binding Site Improves Rigid Local Antibody-Antigen Docking." *Protein Engineering, Design & Selection : PEDS* 26(10):621–29.

Krivov, G. G., Shapovalov, M. V, and Dunbrack, R. L. 2009. "Improved Prediction of Protein Side-Chain Conformations with SCWRL4." *Proteins* 77(4):778–95.

Kuhlman, B., Dantas, G., Ireton, G. C., Varani, G., Stoddard, B. L., and Baker, D. 2003. "Design of a Novel Globular Protein Fold with Atomic-Level Accuracy." *Science* 302(5649):1364–68.

Kunik, V., Ashkenazi, S., and Ofran, Y. 2012. "Paratome: An Online Tool for Systematic Identification of Antigen-Binding Regions in Antibodies Based on Sequence or Structure." *Nucleic Acids Research* 40(Web Server issue):W521-4.

Kunik, V., Peters, B., and Ofran, Y. 2012. "Structural Consensus among Antibodies Defines the Antigen Binding Site." *PLoS Computational Biology* 8(2):e1002388.

Kuroda, D., Shirai, H., Jacobson, M. P., and Nakamura, H. 2012. "Computer-Aided Antibody Design." *Protein Engineering Design and Selection* (25(10)):507–22.

Kuroda, D., Shirai, H., Kobori, M., and Nakamura, H. 2008. "Structural Classification of CDR-H3 Revisited: A Lesson in Antibody Modeling." *Proteins: Structure, Function, and Bioinformatics* 73(3):608–20.

Kuroda, D., Shirai, H., Kobori, M., and Nakamura, H. 2009. "Systematic Classification of CDR-L3 in Antibodies: Implications of the Light Chain Subtypes and the VL--VH Interface." *Proteins: Structure, Function, and Bioinformatics* 75(1):139–46.

Lapidoth, G. D., Baran, D., Pszolla, G. M., Norn, C., Alon, A., Tyka, M. D., and Fleishman, S. J. 2015. "AbDesign: An Algorithm for Combinatorial Backbone Design Guided by Natural Conformations and Sequences." *Proteins* 83(8):1385–1406.

Leaver-Fay, A., Jacak, R., Stranges, P. B., and Kuhlman, B. 2011. "A Generic Program for Multistate Protein Design" edited by V. N. Uversky. *PLoS ONE* 6(7):e20937.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., Renfrew, P. D., Smith, C. A., Sheffler, W., et al. 2011. "ROSETTA3: An Object-Oriented Software Suite for the Simulation and Design of Macromolecules." *Methods in Enzymology* 487:545–74.

Leem, J., Dunbar, J., Georges, G., Shi, J., and Deane, C. M. 2016. "ABodyBuilder: Automated Antibody Structure Prediction with Data–driven Accuracy Estimation." *mAbs* 8(7):1259–68.

Lefranc, M.-P. 2011. "IMGT, the International ImMunoGeneTics Information System." *Cold Spring Harbor Protocols* 2011(6):595–603.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Bodmer, J., Müller, W., Bontrop, R., Lemaitre, M., Malik, A., Barbié, V., and Chaume, D. 1999. "IMGT, the International ImMunoGeneTics Database." *Nucleic Acids Research* 27(1):209–12.

Lefranc, M.-P., Giudicelli, V., Ginestoux, C., Jabado-Michaloud, J., Folch, G., Bellahcene,

F., Wu, Y., Gemrot, E., Brochet, X., Lane, J., et al. 2009. "IMGT, the International ImMunoGeneTics Information System." *Nucleic Acids Research* 37(Database issue):D1006-12.

Lefranc, M.-P., Pommié, C., Ruiz, M., Giudicelli, V., Foulquier, E., Truong, L., Thouvenin-Contet, V., and Lefranc, G. 2003. "IMGT Unique Numbering for Immunoglobulin and T Cell Receptor Variable Domains and Ig Superfamily V-like Domains." *Developmental and Comparative Immunology* 27(1):55–77.

Li, T., Pantazes, R. J., and Maranas, C. D. 2014. "OptMAVEn--a New Framework for the de Novo Design of Antibody Variable Region Models Targeting Specific Antigen Epitopes." *PloS One* 9(8):e105954.

Li, Y., Li, H., Smith-Gill, S. J., and Mariuzza, R. A. 2000. "Three-Dimensional Structures of the Free and Antigen-Bound Fab from Monoclonal Antilysozyme Antibody HyHEL-63." *Biochemistry* 39(21):6296–6309.

Li, Z., Woo, C. J., Iglesias-Ussel, M. D., Ronai, D., and Scharff, M. D. 2004. "The Generation of Antibody Diversity through Somatic Hypermutation and Class Switch Recombination." *Genes & Development* 18(1):1–11.

Liang, S., Zhang, C., and Zhou, Y. 2014. "LEAP: Highly Accurate Prediction of Protein Loop Conformations by Integrating Coarse-Grained Sampling and Optimized Energy Scores with All-Atom Refinement of Backbone and Side Chains." *Journal of*

Computational Chemistry 35(4):335–41.

Lippow, S. M., Wittrup, K. D., and Tidor, B. 2007. "Computational Design of Antibody-Affinity Improvement beyond in Vivo Maturation." *Nature Biotechnology* 25(10):1171–76.

Lonberg, N., Taylor, L. D., Harding, F. A., Trounstine, M., Higgins, K. M., Schramm, S. R., Kuo, C.-C., Mashayekh, R., Wymore, K., McCabe, J. G., et al. 1994. "Antigen-Specific Human Antibodies from Mice Comprising Four Distinct Genetic Modifications." *Nature* 368(6474):856–59.

Looger, L. L., Dwyer, M. A., Smith, J. J., and Hellinga, H. W. 2003. "Computational Design of Receptor and Sensor Proteins with Novel Functions." *Nature* 423(6936):185–90.

Lovell, S. C., Davis, I. W., Arendall, W. B., De Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S., and Richardson, D. C. 2003. "Structure Validation by Cα Geometry: Φ,ψ and Cβ Deviation." *Proteins: Structure, Function and Genetics* 50(3):437–50.

Maaten, L. van der and Hinton, G. 2008. "Visualizing Data Using T-SNE." *Journal of Machine Learning Research 1* 620(1):267–84.

MacKerell, A. D., Bashford, D., Bellott, M., Dunbrack, R. L., Evanseck, J. D., Field, M. J., Fischer, S., Gao, J., Guo, H., Ha, S., et al. 1998. "All-Atom Empirical Potential for

Molecular Modeling and Dynamics Studies of Proteins." *The Journal of Physical Chemistry B* 102(18):3586–3616.

Maggon, K. 2007. "Monoclonal Antibody Gold Rush." *Current Medicinal Chemistry* 14(18):1978–87.

Maloney, D. G., Grillo-López, A. J., White, C. A., Bodkin, D., Schilder, R. J., Neidhart, J. A., Janakiraman, N., Foon, K. A., Liles, T. M., Dallaire, B. K., et al. 1997. "IDEC-C2B8 (Rituximab) Anti-CD20 Monoclonal Antibody Therapy in Patients with Relapsed Low-Grade Non-Hodgkin's Lymphoma." *Blood* 90(6):2188–95.

Manivel, V., Sahoo, N. C., Salunke, D. M., and Rao, K. V. S. 2000. "Maturation of an Antibody Response Is Governed by Modulations in Flexibility of the Antigen-Combining Site." *Immunity* 13(5):611–20.

Marks, C., Deane, C., and Shi, J. 2016. "Hybrid Methods for Protein Loop Modelling." University of Oxford.

Marks, C., Nowak, J., Klostermann, S., Georges, G., Dunbar, J., Shi, J., Kelm, S., and Deane, C. M. 2017. "Sphinx: Merging Knowledge-Based and Ab Initio Approaches to Improve Protein Loop Prediction." *Bioinformatics* 33(9):1346–53.

Martin, A. C. R. 2001. "Protein Sequence and Structure Analysis of Antibody Variable Domains." 422–42.

Martin, A. C. R. and Thornton, J. M. 1996. "Structural Families in Loops of Homologous Proteins: Automatic Classification, Modelling and Application to Antibodies." *Journal of Molecular Biology* 263(5):800–815.

Marze, N. A., Lyskov, S., and Gray, J. J. 2016. "Improved Prediction of Antibody V_L–V_H Orientation." *Protein Engineering Design and Selection* gzw013.

McCafferty, J., Griffiths, A. D., Winter, G., and Chiswell, D. J. 1990. "Phage Antibodies: Filamentous Phage Displaying Antibody Variable Domains." *Nature* 348(6301):552–54.

Mease, P. J. 2007. "Adalimumab in the Treatment of Arthritis." *Therapeutics and Clinical Risk Management* 3(1):133–48.

Méndez, R., Leplae, R., De Maria, L., and Wodak, S. J. 2003. "Assessment of Blind Predictions of Protein-Protein Interactions: Current Status of Docking Methods." *Proteins: Structure, Function and Genetics* 52(1):51–67.

Messih, M. A., Lepore, R., and Tramontano, A. 2015. "LoopIng: A Template-Based Tool for Predicting the Structure of Protein Loops." *Bioinformatics (Oxford, England)* 31(23):3767–72.

Midelfort, K. S., Hernandez, H. H., Lippow, S. M., Tidor, B., Drennan, C. L., and Wittrup, K. D. 2004. "Substantial Energetic Improvement with Minimal Structural Perturbation

in a High Affinity Mutant Antibody." *Journal of Molecular Biology* 343(3):685–701.

Mintseris, J., Pierce, B., Wiehe, K., Anderson, R., Chen, R., and Weng, Z. 2007. "Integrating Statistical Pair Potentials into Protein Complex Prediction." *Proteins: Structure, Function and Genetics* 69(3):511–20.

Mirsky, A., Kazandjian, L., and Anisimova, M. 2015. "Antibody-Specific Model of Amino Acid Substitution for Immunological Inferences from Alignments of Antibody Sequences." *Molecular Biology and Evolution* 32(3):806–19.

Morcos, F., Pagnani, A., Lunt, B., Bertolino, A., Marks, D. S., Sander, C., Zecchina, R., Onuchic, J. N., Hwa, T., and Weigt, M. 2011. "Direct-Coupling Analysis of Residue Coevolution Captures Native Contacts across Many Protein Families." *Proceedings of the National Academy of Sciences of the United States of America* 108(49):E1293-301.

Morea, V., Lesk, A. M., and Tramontano, A. 2000. "Antibody Modeling: Implications for Engineering and Design." *Methods* 20(3):267–79.

Morea, V., Tramontano, A., Rustici, M., Chothia, C., and Lesk, A. M. 1997. "Antibody Structure, Prediction and Redesign." *Biophysical Chemistry* 68(1):9–16.

Morea, V., Tramontano, A., Rustici, M., Chothia, C., and Lesk, A. M. 1998. "Conformations of the Third Hypervariable Region in the VH Domain of Immunoglobulins." *Journal of*

Molecular Biology 275(2):269–94.

Morrison, S. L., Johnson, M. J., Herzenberg, L. A., and Oi, V. T. 1984. "Chimeric Human Antibody Molecules: Mouse Antigen-Binding Domains with Human Constant Region Domains." *Proceedings of the National Academy of Sciences of the United States of America* 81(21):6851–55.

Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T., and Tramontano, A. 2014. "Critical Assessment of Methods of Protein Structure Prediction (CASP) - Round X." *Proteins: Structure, Function, and Bioinformatics* 82(S2):1–6.

Mukherjee, S. and Zhang, Y. 2009. "MM-Align: A Quick Algorithm for Aligning Multiple-Chain Protein Complex Structures Using Iterative Dynamic Programming." *Nucleic Acids Research* 37(11):e83–e83.

Murphy, K., Travers, P., Walport, M., and Janeway, C. 2012. *Janeway's Immunobiology*. 8th ed. New York, New York, USA: Garland Science.

Narayanan, A., Sellers, B. D., and Jacobson, M. P. 2009. "Energy-Based Analysis and Prediction of the Orientation between Light- and Heavy-Chain Antibody Variable Domains." *Journal of Molecular Biology* 388(5):941–53.

Needleman, S. B. and Wunsch, C. D. 1970. "A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins." *Journal of Molecular*

Biology 48(3):443–53.

Nikoloudis, D., Pitts, J. E., and Saldanha, J. W. 2014. "A Complete, Multi-Level Conformational Clustering of Antibody Complementarity-Determining Regions." *PeerJ* 2:e456.

Nilmeier, J., Hua, L., Coutsias, E. A., and Jacobson, M. P. 2011. "Assessing Protein Loop Flexibility by Hierarchical Monte Carlo Sampling." *Journal of Chemical Theory and Computation* 7(5):1564–74.

North, B., Lehmann, A., and Dunbrack Jr, R. L. 2011. "A New Clustering of Antibody CDR Loop Conformations." *Journal of Molecular Biology* 406(2):228–56.

Nowak, J., Baker, T., Georges, G., Kelm, S., Klostermann, S., Shi, J., Sridharan, S., and Deane, C. M. 2016. "Length-Independent Structural Similarities Enrich the Antibody CDR Canonical Class Model." *mAbs* 8(4):751–60.

O'Meara, M. J., Leaver-Fay, A., Tyka, M. D., Stein, A., Houlihan, K., DiMaio, F., Bradley, P., Kortemme, T., Baker, D., Snoeyink, J., et al. 2015. "Combined Covalent-Electrostatic Model of Hydrogen Bonding Improves Structure Prediction with Rosetta." *Journal of Chemical Theory and Computation* 11(2):609–22.

Oliva, B., Bates, P. A., Querol, E., Avilés, F. X., and Sternberg, M. J. 1998. "Automated Classification of Antibody Complementarity Determining Region 3 of the Heavy

Chain (H3) Loops into Canonical Forms and Its Application to Protein Structure Prediction." *Journal of Molecular Biology* 279(5):1193–1210.

Olsson, T. S. G., Williams, M. A., Pitt, W. R., and Ladbury, J. E. 2008. "The Thermodynamics of Protein-Ligand Interaction and Solvation: Insights for Ligand Design." *Journal of Molecular Biology* 384(4):1002–17.

Pantazes, R. J. and Maranas, C. D. 2010. "OptCDR: A General Computational Method for the Design of Antibody Complementarity Determining Regions for Targeted Epitope Binding." *Protein Engineering, Design & Selection : PEDS* 23(11):849–58.

Parks, T., Mirabel, M. M., Kado, J., Auckland, K., Nowak, J., Rautanen, A., Mentzer, A. J., Marijon, E., Jouven, X., Perman, M. L., et al. 2017. "Association between a Common Immunoglobulin Heavy Chain Allele and Rheumatic Heart Disease Risk in Oceania." *Nature Communications* 8:14946.

Pauling, L., Corey, R. B., and Branson, H. R. 1951. "The Structure of Proteins; Two Hydrogen-Bonded Helical Configurations of the Polypeptide Chain." *Proceedings of the National Academy of Sciences of the United States of America* 37(4):205–11.

Pearson, K. 1901. "LIII. On Lines and Planes of Closest Fit to Systems of Points in Space." *Philosophical Magazine Series 6* 2(11):559–72.

Pierce, B. G., Hourai, Y., Weng, Z., Vajda, S., and Jaroszewski, L. 2011. "Accelerating

Protein Docking in ZDOCK Using an Advanced 3D Convolution Library." *PLoS ONE* 6(9):e24657.

Poosarla, V. G., Li, T., Goh, B. C., Schulten, K., Wood, T. K., and Maranas, C. D. 2017. "Computational de Novo Design of Antibodies Binding to a Peptide with High Affinity." *Biotechnology and Bioengineering* 114(6):1331–42.

Qian, N. and Sejnowski, T. J. 1988. "Predicting the Secondary Structure of Globular Proteins Using Neural Network Models." *Journal of Molecular Biology* 202(4):865–84.

Queen, C., Schneider, W. P., Selick, H. E., Payne, P. W., Landolfi, N. F., Duncan, J. F., Avdalovic, N. M., Levitt, M., Junghans, R. P., and Waldmann, T. A. 1989. "A Humanized Antibody That Binds to the Interleukin 2 Receptor." *Proceedings of the National Academy of Sciences of the United States of America* 86(24):10029–33.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. 1963. "Stereochemistry of Polypeptide Chain Configurations." *Journal of Molecular Biology* 7(1):95–99.

Reczko, M., Martin, A. C. R., Bohr, H., and Suhai, S. 1995. "Prediction of Hypervariable CDR-H3 Loop Structures in Antibodies." *Protein Engineering* 8(4):389–95.

Rees, A. R., Staunton, D., Webster, D. M., Searle, S. J., Henry, A. H., and Pedersen, J. T. 1994. "Antibody Design: Beyond the Natural Limits." *Trends in Biotechnology*

12(5):199–206.

Regep, C., Georges, G., Shi, J., Popovic, B., and Deane, C. M. 2017. "The H3 Loop of Antibodies Shows Unique Structural Characteristics." *Proteins: Structure, Function and Bioinformatics* 85(7):1311–18.

Reid, C., Rushe, M., Jarpe, M., van Vlijmen, H., Dolinski, B., Qian, F., Cachero, T. G., Cuervo, H., Yanachkova, M., Nwankwo, C., et al. 2006. "Structure Activity Relationships of Monocyte Chemoattractant Proteins in Complex with a Blocking Antibody." *Protein Engineering, Design & Selection : PEDS* 19(7):317–24.

Retter, I., Althaus, H. H., Münch, R., and Müller, W. 2005. "VBASE2, an Integrative V Gene Database." *Nucleic Acids Research* 33(Database issue):D671-4.

Richardson, J. S. 1981. "The Anatomy and Taxonomy of Protein Structure." *Advances in Protein Chemistry* 34(C):167–339.

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., and Baker, D. 2004. "Protein Structure Prediction Using Rosetta." *Methods in Enzymology* 383:66–93.

Samudrala, R. and Moult, J. 1998. "An All-Atom Distance-Dependent Conditional Probability Discriminatory Function for Protein Structure Prediction." *Journal of Molecular Biology* 275(5):895–916.

Sato, J. D., Kawamoto, T., Le, A. D., Mendelsohn, J., Polikoff, J., and Sato, G. H. 1983.

"Biological Effects in Vitro of Monoclonal Antibodies to Human Epidermal Growth Factor Receptors." *Molecular Biology & Medicine* 1(5):511–29.

Sau, N. Van, Cuong, D. C., Quang, L. S., and Vinh, L. S. 2011. "Protein Type Specific Amino Acid Substitution Models for Influenza Viruses." Pp. 98–103 in *2011 Third International Conference on Knowledge and Systems Engineering*. IEEE.

Scharf, L., Scheid, J. F., Lee, J. H., West, A. P., Chen, C., Gao, H., Gnanapragasam, P. N. P., Mares, R., Seaman, M. S., Ward, A. B., et al. 2014. "Antibody 8ANC195 Reveals a Site of Broad Vulnerability on the HIV-1 Envelope Spike." *Cell Reports* 7(3):785–95.

Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., and Wolfson, H. J. 2005. "PatchDock and SymmDock: Servers for Rigid and Symmetric Docking." *Nucleic Acids Research* 33(Web Server):W363–67.

Schölkopf, B., Smola, A. J., and Muller, K. R. 1999. "Kernel Principal Component Analysis." *Advances in Kernel Methods Support Vector Learning* 1327(3):327–52.

Schrodinger LLC. 2010. "The PyMOL Molecular Graphics System, Version 1.7.4."

Sela-Culang, I., Alon, S., and Ofran, Y. 2012. "A Systematic Comparison of Free and Bound Antibodies Reveals Binding-Related Conformational Changes." *The Journal of Immunology* 189(10):4890–99.

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

Senin, P. 2008. "Dynamic Time Warping Algorithm Review." *Information and Computer Science Department University of Hawaii at Manoa Honolulu, USA* 1–23.

Sharp, K. A. and Honig, B. 1990. "Calculating Total Electrostatic Energies with the Nonlinear Poisson-Boltzmann Equation." *The Journal of Physical Chemistry* 94(19):7684–92.

Sheffler, W. and Baker, D. 2009. "RosettaHoles: Rapid Assessment of Protein Core Packing for Structure Prediction, Refinement, Design, and Validation." *Protein Science : A Publication of the Protein Society* 18(1):229–39.

Shen, M.-Y. and Sali, A. 2006. "Statistical Potential for Assessment and Prediction of Protein Structures." *Protein Science : A Publication of the Protein Society* 15(11):2507–24.

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

Shirai, H., Kidera, A., and Nakamura, H. 1999. "H3-Rules: Identification of CDR-H3 Structures in Antibodies." *FEBS Letters* 455(1):188–97.

Shlomchik, M. J. and Weisel, F. 2012. "Germinal Center Selection and the Development

of Memory B and Plasma Cells." *Immunological Reviews* 247(1):52–63.

Sircar, A. and Gray, J. J. 2010. "SnugDock: Paratope Structural Optimization during Antibody-Antigen Docking Compensates for Errors in Antibody Homology Models." *PLoS Computational Biology* 6(1):e1000644.

Sivasubramanian, A., Sircar, A., Chaudhury, S., and Gray, J. J. 2009. "Toward High-Resolution Homology Modeling of Antibody Fv Regions and Application to Antibody-Antigen Docking." *Proteins* 74(2):497–514.

Smith, C. A. and Kortemme, T. 2008. "Backrub-like Backbone Simulation Recapitulates Natural Protein Conformational Variability and Improves Mutant Side-Chain Prediction." *Journal of Molecular Biology* 380(4):742–56.

Sokal, R. R. and Michener, C. D. 1958. "A Statistical Method for Evaluating Systematic Relationships." *University of Kansas Science Bulletin* 38:1409–38.

Soto, C. S., Fasnacht, M., Zhu, J., Forrest, L., and Honig, B. 2008. "Loop Modeling: Sampling, Filtering, and Scoring." *Proteins* 70(3):834–43.

Srivastava, N., Hinton, G. E., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. 2014. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." *Journal of Machine Learning Research (JMLR)* 15:1929–58.

Stein, A., Kortemme, T., Fleishman, S., Baker, D., Sohl, J., Jaswal, S., Agard, D., Das, R.,

Mandell, D., Coutsias, E., et al. 2013. "Improvements to Robotics-Inspired Conformational Sampling in Rosetta" edited by Y. Zhang. *PLoS ONE* 8(5):e63090.

Stranges, P. B. and Kuhlman, B. 2013. "A Comparison of Successful and Failed Protein Interface Designs Highlights the Challenges of Designing Buried Hydrogen Bonds." *Protein Science : A Publication of the Protein Society* 22(1):74–82.

Tan, C.-W., Jones, D. T., Altschul, S., Madden, T., Schaffer, A., Zhang, J., Zhang, Z., Miller, W., Lipman, D., Moult, J., et al. 2008. "Using Neural Networks and Evolutionary Information in Decoy Discrimination for Protein Tertiary Structure Prediction." *BMC Bioinformatics* 9(1):94.

Tax, N., Bockting, S., and Hiemstra, D. 2015. "A Cross-Benchmark Comparison of 87 Learning to Rank Methods." *Information Processing and Management* 51(6):757–72.

Teng, G. and Papavasiliou, F. N. 2007. "Immunoglobulin Somatic Hypermutation." *Annual Review of Genetics* 41(1):107–20.

Teplyakov, A. and Gilliland, G. L. 2014. "Canonical Structures of Short CDR-L3 in Antibodies." *Proteins: Structure, Function, and Bioinformatics* 82(8):1668–73.

Teplyakov, A., Luo, J., Obmolova, G., Malia, T. J., Sweet, R., Stanfield, R. L., Kodangattil, S., Almagro, J. C., and Gilliland, G. L. 2014. "Antibody Modeling Assessment II.

Structures and Models." *Proteins* 82(8):1563–82.

Teplyakov, A., Obmolova, G., Carton, J. M., Gao, W., Zhao, Y., and Gilliland, G. L. 2010. "On the Domain Pairing in Chimeric Antibodies." *Molecular Immunology* 47(14):2422–26.

Teplyakov, A., Obmolova, G., Malia, T. J., Luo, J., Muzammil, S., Sweet, R., Almagro, J. C., and Gilliland, G. L. 2016. "Structural Diversity in a Human Antibody Germline Library." *mAbs* 8(6):1045–63.

Tharakaraman, K., Robinson, L. N., Hatas, A., Chen, Y.-L., Siyue, L., Raguram, S., Sasisekharan, V., Wogan, G. N., and Sasisekharan, R. 2013. "Redesign of a Cross-Reactive Antibody to Dengue Virus with Broad-Spectrum Activity and Increased in Vivo Potency." *Proceedings of the National Academy of Sciences of the United States of America* 110(17):E1555-64.

The UniProt Consortium. 2017. "UniProt: The Universal Protein Knowledgebase." *Nucleic Acids Research* 45(D1):D158–69.

Tomlinson, I. M., Cox, J. P., Gherardi, E., Lesk, A. M., and Chothia, C. 1995. "The Structural Repertoire of the Human V Kappa Domain." *The EMBO Journal* 14(18):4628–4638.

Tonegawa, S. 1983. "Somatic Generation of Antibody Diversity." *Nature*

302(5909):575–81.

Tovchigrechko, A., Wells, C. A., and Vakser, I. A. 2002. "Docking of Protein Models." *Protein Science* 11(8):1888–96.

Tramontano, A., Chothia, C., and Lesk, A. M. 1990. "Framework Residue 71 Is a Major Determinant of the Position and Conformation of the Second Hypervariable Region in the VH Domains of Immunoglobulins." *Journal of Molecular Biology* 215(1):175–82.

Tran, K., Poulsen, C., Guenaga, J., de Val, N., Wilson, R., Sundling, C., Li, Y., Stanfield, R. L., Wilson, I. A., Ward, A. B., et al. 2014. "Vaccine-Elicited Primate Antibodies Use a Distinct Approach to the HIV-1 Primary Receptor Binding Site Informing Vaccine Redesign." *Proceedings of the National Academy of Sciences* 111(7):E738–47.

Walsh, I., Martin, A. J., Mooney, C., Rubagotti, E., Vullo, A., and Pollastri, G. 2009. "Ab Initio and Homology Based Prediction of Protein Domains by Recursive Neural Networks." *BMC Bioinformatics* 10(1):195.

Wang, F., Ekiert, D. C., Ahmad, I., Yu, W., Zhang, Y., Bazirgan, O., Torkamani, A., Raudsepp, T., Mwangi, W., Criscitiello, M. F., et al. 2013. "Reshaping Antibody Diversity." *Cell* 153(6):1379–93.

Wang, G. and Dunbrack, R. L. 2003. "PISCES: A Protein Sequence Culling Server."

Bioinformatics (Oxford, England) 19(12):1589–91.

Wang, Y., Wang, L., Li, Y., He, D., Liu, T.-Y., and Chen, W. 2013. "A Theoretical Analysis of NDCG Type Ranking Measures."

Ward, J. H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association* 58(301):236–44.

Weikl, T. R. and von Deuster, C. 2009. "Selected-Fit versus Induced-Fit Protein Binding: Kinetic Differences and Mutational Analysis." *Proteins: Structure, Function, and Bioinformatics* 75(1):104–10.

Weitzner, B. D., Dunbrack Jr, R. L., and Gray, J. J. 2015. "The Origin of CDR H3 Structural Diversity." *Structure* 23(2):302–11.

Weitzner, B. D., Jeliazkov, J. R., Lyskov, S., Marze, N., Kuroda, D., Frick, R., Adolf-Bryfogle, J., Biswas, N., Dunbrack, R. L., and Gray, J. J. 2017. "Modeling and Docking of Antibody Structures with Rosetta." *Nature Protocols* 12(2):401–16.

Weitzner, B. D., Kuroda, D., Marze, N., Xu, J., and Gray, J. J. 2014. "Blind Prediction Performance of RosettaAntibody 3.0: Grafting, Relaxation, Kinematic Loop Modeling, and Full CDR Optimization." *Proteins* 82(8):1611–23.

Wilson, P. C., de Bouteiller, O., Liu, Y. J., Potter, K., Banchereau, J., Capra, J. D., and Pascual, V. 1998. "Somatic Hypermutation Introduces Insertions and Deletions into

Immunoglobulin V Genes." *The Journal of Experimental Medicine* 187(1):59–70.

Wu, T. T. and Kabat, E. A. 1970. "An Analysis of the Sequences of the Variable Regions of Bence Jones Proteins and Myeloma Light Chains and Their Implications for Antibody Complementarity." *The Journal of Experimental Medicine* 132(2):211–50.

Wu, X., Zhou, T., Zhu, J., Zhang, B., Georgiev, I., Wang, C., Chen, X., Longo, N. S., Louder, M., McKee, K., et al. 2011. "Focused Evolution of HIV-1 Neutralizing Antibodies Revealed by Structures and Deep Sequencing." *Science* 333(6049):1593–1602.

Xiang, Z., Soto, C. S., and Honig, B. 2002. "Evaluating Conformational Free Energies: The Colony Energy and Its Application to the Problem of Loop Prediction." *Proceedings of the National Academy of Sciences* 99(11):7432–37.

Xu, J., Tack, D., Hughes, R. A., Ellington, A. D., and Gray, J. J. 2014. "Structure-Based Non-Canonical Amino Acid Design to Covalently Crosslink an Antibody-Antigen Complex." *Journal of Structural Biology* 185(2):215–22.

Xu, J. and Zhang, Y. 2010. "How Significant Is a Protein Structure Similarity with TM-Score = 0.5?" *Bioinformatics* 26(7):889–95.

Ye, J., Ma, N., Madden, T. L., and Ostell, J. M. 2013. "IgBLAST: An Immunoglobulin Variable Domain Sequence Analysis Tool." *Nucleic Acids Research* 41(Web Server issue):W34-40.

Zdanov, A., Li, Y., Bundle, D. R., Deng, S. J., MacKenzie, C. R., Narang, S. A., Young, N. M., and Cygler, M. 1994. "Structure of a Single-Chain Antibody Variable Domain (Fv) Fragment Complexed with a Carbohydrate Antigen at 1.7A Resolution." *Proceedings of the National Academy of Sciences of the United States of America* 91(14):6423–27.

Zhang, J. and Zhang, Y. 2010. "A Novel Side-Chain Orientation Dependent Potential Derived from Random-Walk Reference State for Protein Fold Selection and Structure Prediction." *PloS One* 5(10):e15386.

Zhang, Y. and Skolnick, J. 2004. "Scoring Function for Automated Assessment of Protein Structure Template Quality." *Proteins* 57(4):702–10.

Zhang, Y. and Skolnick, J. 2005. "TM-Align: A Protein Structure Alignment Algorithm Based on the TM-Score." *Nucleic Acids Research* 33(7):2302–9.

Zhou, H. and Skolnick, J. 2011. "GOAP: A Generalized Orientation-Dependent, All-Atom Statistical Potential for Protein Structure Prediction." *Biophysical Journal* 101(8):2043–52.

Zhou, H. and Zhou, Y. 2002. "Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction." *Protein Science : A Publication of the Protein Society* 11(11):2714–26.

8 APPENDICES